Online-to-PAC Conversions:
Generalization Bounds via Regret Analysis

Gergely Neu

upf. Universitat Pompeu Fabra Barcelona

joint work with Gábor Lugosi

Funded by
ERC StG
ScaleR

The plan for today

- Statistical learning crash course
- Online learning crash course
- From regret analysis to generalization bounds
- Some examples

The plan for today

- Statistical learning crash course
- Online learning crash course
- From regret analysis to generalization bounds
- Some examples

~

We will construct online learning algorithms that will certify bounds on the generalization error of a given statistical learning algorithm.

~

The plan for today

- Statistical learning crash course
- Online learning crash course
- From regret analysis to generalization bounds
- Some examples

~

We will construct online learning algorithms that will certify bounds on the generalization error of a given statistical learning algorithm.

~

Setup: Statistical 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: Statistical 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 $W_n = \mathcal{A}(S_n)$

Goal:

understand when algorithm $\mathcal{A}$ produces W_n with small risk $R(W_n) = \mathbb{E}_{Z'}\left[\ell(W_n, Z')|W_n\right]$

Risk vs. empirical risk

- Risk: $R(w) = \mathbb{E}_Z[\ell(w, Z)]$
- Empirical risk: $\hat{R}(w, S_n) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i)$
- Risk decomposition for $W_n = \mathcal{A}(S_n)$:

$$R(W_n) = \hat{R}(W_n, S_n) + \left(R(W_n) - \hat{R}(W_n, S_n)\right)$$

Risk vs. empirical risk

- Risk: $R(w) = \mathbb{E}_Z[\ell(w, Z)]$
- Empirical risk: $\hat{R}(w, S_n) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i)$
- Risk decomposition for $W_n = \mathcal{A}(S_n)$:

$$R(W_n) = \hat{R}(W_n, S_n) + \underbrace{\left(R(W_n) - \hat{R}(W_n, S_n)\right)}_{\substack{\text{generalization error} \\ \operatorname{gen}(W_n, S_n)}}$$

Risk vs. empirical risk

- Risk: $R(w) = \mathbb{E}_Z[\ell(w, Z)]$
- Empirical risk: $\hat{R}(w, S_n) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i)$
- Risk decomposition for $W_n = \mathcal{A}(S_n)$:

$$R(W_n) = \hat{R}(W_n, S_n) + \underbrace{\left(R(W_n) - \hat{R}(W_n, S_n)\right)}_{\substack{\text{generalization error} \\ \mathrm{gen}(W_n, S_n)}}$$

Directly controlled by algorithm

Risk vs. empirical risk

- Risk: $R(w) = \mathbb{E}_Z[\ell(w,Z)]$
- Empirical risk: $\hat{R}(w,S_n) = \frac{1}{n}\sum_{i=1}^{n} \ell(w,Z_i)$
- Risk decomposition for $W_n = \mathcal{A}(S_n)$:

$$R(W_n) = \hat{R}(W_n,S_n) + \underbrace{\Big(R(W_n) - \hat{R}(W_n,S_n)\Big)}_{\substack{\text{generalization error}\\ \mathrm{gen}(W_n,S_n)}}$$

Directly controlled by algorithm

The BIG question:

why/when is this small?

Analyzing the generalization error

- Uniform convergence: bound $\sup_w \left| R(w) - \hat{R}(w, S_n) \right|$
  - Distribution-agnostic: VC-dimension
  - Distribution-dependent: Rademacher complexity, margin conditions

Analyzing the generalization error

- Uniform convergence: bound $\sup_w \left| R(w) - \hat{R}(w, S_n) \right|$
  - Distribution-agnostic: VC-dimension
  - Distribution-dependent: Rademacher complexity, margin conditions

- Algorithm-dependent:
  - Stability (Bousquet & Eliseeff, 2002)
  - PAC-Bayes (Shawe-Taylor & Williamson, 1997, McAllester, 1998, Langford and Seeger, 2001)
  - Information-theoretic (Russo & Zou, 2016, Xu & Raginsky, 2017)

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 $\mathcal{A}$,

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

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 $\mathcal{A}$,

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

Mutual information between W_n and S_n

PAC-Bayes

Theorem

(McAllester, Catoni, Langford, Seeger, etc.)

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

Then, for any prior $P_0 \in \Delta_{\mathcal{W}}$, w.p. $\geq 1 - \delta$

the following holds for any learning algorithm $\mathcal{A}$:

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)|S_n]| \leq \sqrt{\frac{2\sigma^2 \mathcal{D}_{\mathrm{KL}}\left(P_{W_n|S_n} \middle| P_0\right)}{n}} + \sqrt{\frac{\sigma^2 \log(\log n\,/\delta)}{n}}$$

The plan for today

- Statistical learning crash course
- Online learning crash course
- From regret analysis to generalization bounds
- Some examples

~

We will construct online learning algorithms that will certify bounds on the generalization error of a given statistical learning algorithm.

~

Online learning

The protocol of Online Linear Optimization (OLO)

For each $t = 1,2,\ldots,T$, **repeat**

- Online learner picks decision $P_t \in \mathcal{P}$
- Environment / adversary picks cost function $c_t \in \mathcal{C}$
- Online learner incurs cost $\langle P_t, c_t \rangle$
- Online learner observes cost function c_t

- $\mathcal{P}$ and $\mathcal{C}$ are convex sets in appropriate Banach spaces
- Environment can use all info from the past and even knowledge of the online learner's algorithm

Regret analysis

Performance of the online learner is measured by its regret:

$$\mathfrak{R}_T(P^*) = \sum_{t=1}^{T} \langle P_t, c_t \rangle - \sum_{t=1}^{T} \langle P^*, c_t \rangle$$

Regret analysis

Performance of the online learner is measured by its regret:

$$\mathfrak{R}_T(P^*) = \sum_{t=1}^{T} \langle P_t, c_t \rangle - \sum_{t=1}^{T} \langle P^*, c_t \rangle$$

total cost of online learner

total cost of a fixed comparator $P^* \in \mathcal{P}$

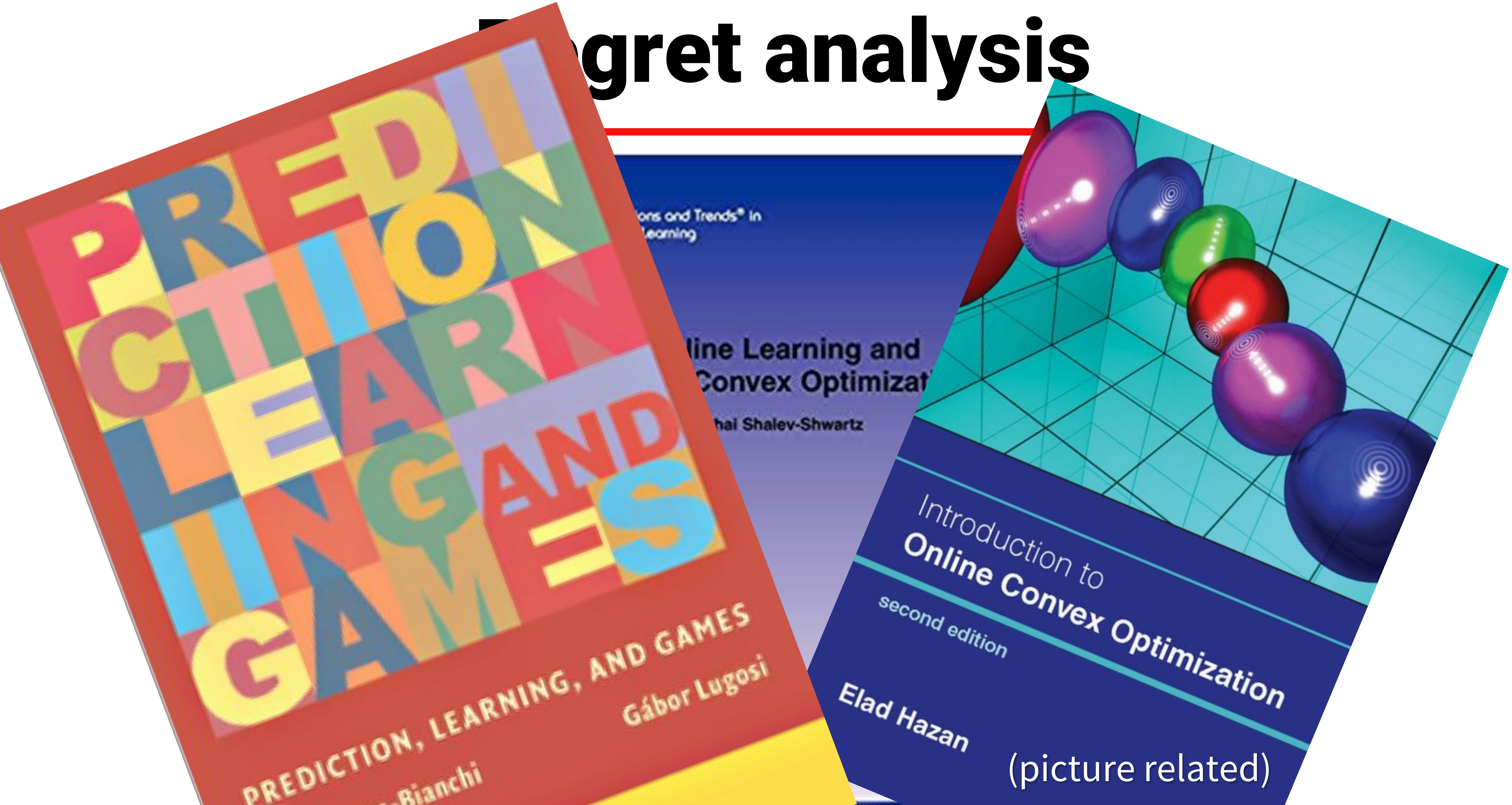
Regret analysis

Performance of the online learner is measured by its regret:

$$\mathfrak{R}_T(P^*) = \underbrace{\sum_{t=1}^{T}\langle P_t, c_t\rangle}_{\text{total cost of online learner}} - \underbrace{\sum_{t=1}^{T}\langle P^*, c_t\rangle}_{\text{total cost of a fixed comparator } P^* \in \mathcal{P}}$$

How can we possibly bound this?

Regret analysis



(picture related)

A classic online learning result

- Let $\mathcal{P} = \Delta_{\mathcal{W}}$ be a probability simplex and $\mathcal{C} \in [-\sigma, \sigma]^{\mathcal{W}}$
- Cost is defined as $\langle P, c \rangle = \mathbb{E}_{W \sim P}[c(W)]$

Theorem

(Vovk 1990, Littlestone & Warmuth 1994, Freund & Schapire 1997)

The Exponentially Weighted Averaging algorithm that predicts $P_{t+1}(w) \propto P_t(w) e^{-\eta c_t(w)}$ satisfies the following regret bound:

$$\mathfrak{R}_T(P^*) \leq \frac{\mathcal{D}_{KL}(P^* | P_1)}{\eta} + \frac{\eta \sigma^2 T}{2}$$

A classic online learning result

- Let $\mathcal{P} = \Delta_{\mathcal{W}}$ be a probability simplex and $\mathcal{C} \in [-\sigma, \sigma]^{\mathcal{W}}$
- Cost is defined as $\langle P, c \rangle = \mathbb{E}_{W \sim P}[c(W)]$

Theorem

(Vovk 1990, Littlestone & Warmuth 1994, Freund & Schapire 1997)

The Exponentially Weighted Averaging algorithm that predicts $P_{t+1}(w) \propto P_t(w) e^{-\eta c_t(w)}$ satisfies the following regret bound:

$$\mathfrak{R}_T(P^*) \leq \boxed{\sqrt{T\sigma^2 \mathcal{D}_{KL}(P^* | P_1)}}$$

The plan for today

- Statistical learning crash course
- Online learning crash course
- From regret analysis to generalization bounds
- Some examples

~

We will construct online learning algorithms that will certify bounds on the generalization error of a given statistical learning algorithm.

~

Reduction to online learning

The generalization game

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

- Online learner picks $P_t = \mathrm{Law}(\widetilde{W}_t) \in \Delta_{\mathcal{W}}$
- Environment picks cost function $c_t(w) = \ell(w, Z_t) - \mathbb{E}_{Z'}[\ell(w, Z')]$
- Online learner incurs cost $\langle P_t, c_t \rangle = \mathbb{E}_{\widetilde{W}_t \sim P_t}\left[c_t(\widetilde{W}_t)\right]$
- Online learner observes cost function c_t

Reduction to online learning

The generalization game

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

- Online learner picks $P_t = \mathrm{Law}(\widetilde{W}_t) \ \in \Delta_{\mathcal{W}}$
- Environment picks cost function $c_t(w) = \ell(w, Z_t) - \mathbb{E}_{Z'}[\ell(w, Z')]$
- Online learner incurs cost $\langle P_t, c_t \rangle = \mathbb{E}_{\widetilde{W}_t \sim P_t}\left[c_t(\widetilde{W}_t)\right]$
- Online learner observes cost function c_t

Fits into online learning framework with $T = n, \mathcal{P} = \Delta_{\mathcal{W}}$.
The costs are i.i.d. and zero-mean for any fixed w.

Let’s do some math

- Generalization error can be written as follows:

$$\mathbb{E}[\mathrm{gen}(W_n, S_n)|S_n] = \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\Big[\big(\mathbb{E}_{Z'}[\ell(W_n, Z')] - \ell(W_n, Z_t)\big)\Big|S_n\Big]$$

Let’s do some math

- Generalization error can be written as follows:

$$\mathbb{E}[\text{gen}(W_n, S_n)|S_n] = \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\left[\left(\mathbb{E}_{Z'}[\ell(W_n, Z')] - \ell(W_n, Z_t)\right)\middle|S_n\right]$$
$$= -\frac{1}{n}\sum_{t=1}^{n}\left\langle P_{W_n|S_n}, c_t\right\rangle$$

Let’s do some math

- Generalization error can be written as follows:

$$\begin{aligned}\mathbb{E}[\mathrm{gen}(W_n, S_n)|S_n] &= \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\Big[\big(\mathbb{E}_{Z'}[\ell(W_n, Z')] - \ell(W_n, Z_t)\big)\Big|S_n\Big] \\ &= -\frac{1}{n}\sum_{t=1}^{n} \langle P_{W_n|S_n}, c_t\rangle \\ &= \frac{1}{n}\sum_{t=1}^{n} \langle P_t - P_{W_n|S_n}, c_t\rangle - \frac{1}{n}\sum_{t=1}^{n} \langle P_t, c_t\rangle\end{aligned}$$

Let’s do some math

- Generalization error can be written as follows:

$$\begin{aligned}\mathbb{E}[\mathrm{gen}(W_n, S_n)|S_n] &= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}\Big[\big(\mathbb{E}_{Z'}[\ell(W_n, Z')] - \ell(W_n, Z_t)\big)\Big|S_n\Big] \\ &= -\frac{1}{n}\sum_{t=1}^{n}\langle P_{W_n|S_n}, c_t\rangle \\ &= \underbrace{\frac{1}{n}\sum_{t=1}^{n}\langle P_t - P_{W_n|S_n}, c_t\rangle}_{\text{regret of online learner against comparator } P_{W_n|S_n}} - \underbrace{\frac{1}{n}\sum_{t=1}^{n}\langle P_t, c_t\rangle}_{\text{total cost of online learner}}\end{aligned}$$

regret of online learner against comparator $P_{W_n|S_n}$

total cost of online learner

Magic trick

Lemma

Suppose that the loss function is σ-subgaussian for all w.
Then, with probability $\geq 1-\delta$,

$$\frac{1}{n}\sum_{t=1}^{n}\langle P_t, c_t\rangle \leq \sqrt{\frac{\sigma^2\log(1/\delta)}{2n}}$$

Inspired by

“On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds, and Regularization”

by Kakade, Sridharan, and Tewari (2008)

Proof of magic lemma

- Let's think about the conditional expectation of the cost:

$$\mathbb{E}_t\left[c_t\left(\widetilde{W}_t\right)\right] = \mathbb{E}_t\left[\mathbb{E}_t\left[c_t\left(\widetilde{W}_t\right)\middle|\widetilde{W}_t\right]\right]$$

Proof of magic lemma

- Let’s think about the conditional expectation of the cost:

$$\begin{aligned}\mathbb{E}_t\left[c_t\left(\widetilde{W}_t\right)\right] &= \mathbb{E}_t\left[\mathbb{E}_t\left[c_t\left(\widetilde{W}_t\right)\middle|\widetilde{W}_t\right]\right] \\ &= \mathbb{E}_t\left[\mathbb{E}_t\left[\ell_t\left(\widetilde{W}_t, Z_t\right) - \ell_t\left(\widetilde{W}_t, Z'\right)\middle|\widetilde{W}_t\right]\right]\end{aligned}$$

Proof of magic lemma

- Let’s think about the conditional expectation of the cost:

$$\begin{aligned}\mathbb{E}_t\left[c_t\left(\widetilde{W}_t\right)\right] &= \mathbb{E}_t\left[\mathbb{E}_t\left[c_t\left(\widetilde{W}_t\right)\middle|\widetilde{W}_t\right]\right] \\ &= \mathbb{E}_t\left[\mathbb{E}_t\left[\ell_t\left(\widetilde{W}_t, Z_t\right) - \ell_t\left(\widetilde{W}_t, Z'\right)\middle|\widetilde{W}_t\right]\right] = 0\end{aligned}$$

these two terms are equal because

$\widetilde{W}_t$ is conditionally independent of Z_t:

$$\left(\widetilde{W}_t, Z_t\right)|\mathcal{F}_{t-1} \sim \left(\widetilde{W}_t, Z'\right)|\mathcal{F}_{t-1}$$

Proof of magic lemma

- Let's think about the conditional expectation of the cost:

$$\begin{aligned}\mathbb{E}_t\left[c_t(\widetilde{W}_t)\right] &= \mathbb{E}_t\left[\mathbb{E}_t\left[c_t(\widetilde{W}_t)\middle|\widetilde{W}_t\right]\right]\\ &= \mathbb{E}_t\left[\mathbb{E}_t\left[\ell_t(\widetilde{W}_t, Z_t) - \ell_t(\widetilde{W}_t, Z')\middle|\widetilde{W}_t\right]\right] = 0\end{aligned}$$

these two terms are equal because $\widetilde{W}_t$ is conditionally independent of Z_t:

$$(\widetilde{W}_t, Z_t)|\mathcal{F}_{t-1} \sim (\widetilde{W}_t, Z')|\mathcal{F}_{t-1}$$

$\sum_{t=1}^{n} c_t(\widetilde{W}_t)$ is a martingale, so we can use Azuma—Hoeffding to bound it!!

Proof of magic lemma

- Let’s think about the conditional expectation of the cost:

$$\begin{aligned}\mathbb{E}_t\left[c_t\left(\widetilde{W}_t\right)\right] &= \mathbb{E}_t\left[\mathbb{E}_t\left[c_t\left(\widetilde{W}_t\right)\middle|\widetilde{W}_t\right]\right] \\ &= \mathbb{E}_t\left[\mathbb{E}_t\left[\ell_t\left(\widetilde{W}_t, Z_t\right) - \ell_t\left(\widetilde{W}_t, Z'\right)\middle|\widetilde{W}_t\right]\right] = 0\end{aligned}$$

these two terms are equal because $\widetilde{W}_t$ is conditionally independent of Z_t:

$$\left(\widetilde{W}_t, Z_t\right)|\mathcal{F}_{t-1} \sim \left(\widetilde{W}_t, Z'\right)|\mathcal{F}_{t-1}$$

$\sum_{t=1}^{n} c_t\left(\widetilde{W}_t\right)$ is a martingale, so we can use Azuma—Hoeffding to bound it!!

■

Online-to-PAC conversion

Theorem

Fix an online learning algorithm and let $\mathfrak{R}_n(P^*)$ be its regret against comparator P^*. Suppose that $\mathbb{E}\left[\left(\ell(w,Z)\right)^2\right] \leq V$. Then, with probability at least $1-\delta$, the generalization error of all statistical learning algorithms $W_n = \mathcal{A}(S_n)$ simultaneously satisfy the following bound :

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)|S_n]| \leq \frac{\mathfrak{R}_n\left(P_{W_n|S_n}\right)}{n} + \sqrt{\frac{V\log(1/\delta)}{2n}}$$

Online-to-PAC conversion

Theorem

Fix an online learning algorithm and let $\mathfrak{R}_n(P^*)$ be its regret against comparator P^*. Suppose that $\mathbb{E}\left[\left(\ell(w,Z)\right)^2\right] \le V$. Then, with probability at least $1-\delta$, the generalization error of all statistical learning algorithms $W_n = \mathcal{A}(S_n)$ simultaneously satisfy the following bound :

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)|S_n]| \le \frac{\mathfrak{R}_n\left(P_{W_n|S_n}\right)}{n} + \sqrt{\frac{V\log(1/\delta)}{2n}}$$

the **existence** of an online learning algorithm with bounded regret certifies a bound on the generalization error!!

The plan for today

- Statistical learning crash course
- Online learning crash course
- From regret analysis to generalization bounds
- Some examples

~

We will construct online learning algorithms that will certify bounds on the generalization error of a given statistical learning algorithm.

~

Examples

- PAC-Bayes via Exponential Weighted Averaging
  - McAllester-style bounds
  - Data-dependent bounds
  - Parameter-free bounds
- Generalized PAC-Bayes via Following the Regularized Leader
  - Strongly convex regularizers
  - Empirical bounds via optimistic FTRL
  - Examples: p-norm regularizers, smoothed relative entropy

Examples

- PAC-Bayes via Exponential Weighted Averaging
  - McAllester-style bounds
  - Data-dependent bounds
  - Parameter-free bounds
- Generalized PAC-Bayes via Following the Regularized Leader
  - Strongly convex regularizers
  - Empirical bounds via optimistic FTRL
  - Examples: p-norm regularizers, smoothed relative entropy

PAC-Bayes via EWA

Regret bound of EWA

$$\mathfrak{R}_T(P^*) \leq \frac{\mathcal{D}_{KL}(P^*|P_1)}{\eta} + \frac{\eta\sigma^2 T}{2}$$

$+$

Online-to-PAC

$$\frac{\mathfrak{R}_n\left(P_{W_n|S_n}\right)}{n} + \sqrt{\frac{\sigma^2 \log(1/\delta)}{2n}}$$

PAC-Bayes via EWA

Regret bound of EWA

$$\mathfrak{R}_T(P^*) \leq \frac{\mathcal{D}_{KL}(P^*|P_1)}{\eta} + \frac{\eta\sigma^2 T}{2}$$

$$+$$

Online-to-PAC

$$\frac{\mathfrak{R}_n(P_{W_n|S_n})}{n} + \sqrt{\frac{\sigma^2\log(1/\delta)}{2n}}$$

$$=$$

PAC-Bayes

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)|S_n]| \leq \frac{\mathcal{D}_{KL}(P_{W_n|S_n}|P_1)}{\eta n} + \frac{\eta\sigma^2}{2} + \sqrt{\frac{\sigma^2\log(1/\delta)}{2n}}$$

EWA + steroids

```
Second-order optimistic EWA
Input: learning rate $\eta > 0$, prior $\tilde{P}_1 \in \Delta_{\mathcal{W}}$
Initialization: $C_0 = 0$
For each $t = 1,2, \dots, n$, repeat
    • Calculate $P_t(w) \propto \tilde{P}_t(w) \exp(-\eta g_t(w))$
    • Play action $P_t$, incur cost $\langle P_t, c_t \rangle$, observe $c_t$
    • Calculate auxiliary update
          $\tilde{P}_{t+1}(w) \propto \tilde{P}_t(w) \exp\left(-\eta c_t(w) - \eta^2 \left(c_t(w) - g_t(w)\right)^2\right)$
```

A data-dependent bound

(A regret bound for second-order optimistic EWA)

$+$

Online-to-PAC

$$\frac{\mathfrak{R}_n\left(P_{W_n|S_n}\right)}{n} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

A data-dependent bound

(A regret bound for second-order optimistic EWA)

$+$

Online-to-PAC

$$\frac{\mathfrak{R}_n\left(P_{W_n|S_n}\right)}{n} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

$=$

Second-order PAC-Bayes

$$\begin{aligned}
&|\mathbb{E}[\mathrm{gen}(W_n, S_n)|S_n]| \\
&\le \frac{\mathcal{D}_{KL}\left(P_{W_n|S_n}|P_1\right)}{\eta n} + \frac{\eta}{n}\sum_{t=1}^{n} \mathbb{E}\left[\left(\ell(W_n, Z_t)\right)^2 \middle| S_n\right] + \frac{\log(1/\delta)}{2\eta n}
\end{aligned}$$

A data-dependent bound

(A regret bound for second-order optimistic EWA)

$+$

Online-to-PAC

$$\frac{\mathfrak{R}_n\left(P_{W_n|S_n}\right)}{n} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Fast rate if training error = 0!!

Se order PAC-Bayes

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)|S_n]|$$

$$\leq \frac{\mathcal{D}_{KL}\left(P_{W_n|S_n}\middle|P_1\right)}{\eta n} + \frac{\eta}{n}\sum_{t=1}^{n}\mathbb{E}\left[\left(\ell(W_n, Z_t)\right)^2\middle|S_n\right] + \frac{\log(1/\delta)}{2\eta n}$$

A parameter-free PAC-Bayes bound

Regret of "coin-betting"

$$\mathfrak{R}_T(P^*) \leq \sqrt{3T\mathcal{D}_{KL}(P^*|P_1) + 9T}$$

$+$

Online-to-PAC

$$\frac{\mathfrak{R}_n\left(P_{W_n|S_n}\right)}{n} + \sqrt{\frac{\sigma^2 \log(1/\delta)}{2n}}$$

Orabona and Pál (2016)

A parameter-free PAC-Bayes bound

Regret of “coin-betting”

$$\mathfrak{R}_T(P^*) \le \sqrt{3T\mathcal{D}_{KL}(P^*|P_1) + 9T}$$

Orabona and Pál (2016)

$+$

Online-to-PAC

$$\frac{\mathfrak{R}_n\left(P_{W_n|S_n}\right)}{n} + \sqrt{\frac{\sigma^2 \log(1/\delta)}{2n}}$$

$=$

Parameter-free PAC-Bayes

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)|S_n]| \le \sqrt{\frac{3\mathcal{D}_{KL}\left(P_{W_n|S_n}|P_1\right) + 9}{n}} + \sqrt{\frac{\sigma^2 \log(1/\delta)}{2n}}$$

A parameter-free PAC-Bayes bound

Regret of "coin-betting"

$$\mathfrak{R}_T(P^*) \le \sqrt{3T\mathcal{D}_{KL}(P^*|P_1) + 9T}$$

Orabona a

+

Online-to-PAC

$$\frac{\mathfrak{R}_n(P_{W_n|S_n})}{n} + \sqrt{\frac{\sigma^2 \log(1/\delta)}{2n}}$$

Not even a $\log\log n$ factor!

Parameter-free PAC-Bay

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)|S_n]| \le \sqrt{\frac{3\mathcal{D}_{KL}(P_{W_n|S_n}|P_1) + 9}{n}} + \sqrt{\frac{\sigma^2 \log(1/\delta)}{2n}}$$

Examples

- PAC-Bayes via Exponential Weighted Averaging
  - McAllester-style bounds
  - Data-dependent bounds
  - Parameter-free bounds
- Generalized PAC-Bayes via Following the Regularized Leader
  - Strongly convex regularizers
  - Empirical bounds via optimistic FTRL
  - Examples: p-norm regularizers, smoothed relative entropy

Our favorite workhorse: FTRL

```
Follow the regularized leader
Input: regularization function $h: \Delta_{\mathcal{W}} \to \mathbb{R}_+$, learning rate $\eta > 0$
Initialization: $C_0 = 0$
For each $t = 1,2, \dots, T$, repeat
    • Play action
          $P_t = \arg\min_{P \in \Delta_{\mathcal{W}}} \left\{ \langle P, C_{t-1} \rangle + \frac{1}{\eta} h(P) \right\}$
    • Observe cost function $c_t$ and update $C_t = C_{t-1} + c_t$
```

The regret of FTRL

Theorem

Suppose that h is α-strongly convex w.r.t. $\|\cdot\|$.

Then, the regret of FTRL satisfies $\mathfrak{R}_n(P^*) \le \frac{h(P^*)-h(P_1)}{\alpha\eta} + \eta \sum_{t=1}^{T} \|c_t\|_*^2$.

- h is said to be α-strongly convex w.r.t. $\|\cdot\|$ if it satisfies
  $$h(\lambda P + (1-\lambda)P') \le \lambda h(P) + (1-\lambda)h(P') - \frac{\alpha\lambda(1-\lambda)}{2}\|P - P'\|^2$$
- $\|\cdot\|_*$ is the associated dual norm: $\|c\|_* = \sup_{\|P-P'\|\le 1} \langle P - P', c\rangle$

The regret of FTRL

Theorem

Suppose that h is α-strongly convex w.r.t. $\|\cdot\|$.

Then, the regret of FTRL satisfies $\mathfrak{R}_n(P^*) \leq \sqrt{Th(P^*)B^2/\alpha}$.

(if $\max_t \|c_t\|_* \leq B$)

- h is said to be α-strongly convex w.r.t. $\|\cdot\|$ if it satisfies
$$h(\lambda P + (1-\lambda)P') \leq \lambda h(P) + (1-\lambda)h(P') - \frac{\alpha\lambda(1-\lambda)}{2}\|P-P'\|^2$$
- $\|\cdot\|_*$ is the associated dual norm: $\|c\|_* = \sup_{\|P-P'\|\leq 1} \langle P-P', c\rangle$

Generalized PAC-Bayes via FTRL

Regret bound of FTRL

$$\mathfrak{R}_T(P^*) \le \frac{h(P^*) - h(P_1)}{\eta} + \frac{\eta B^2 T}{2\alpha}$$

$$+$$

Online-to-PAC

$$\frac{\mathfrak{R}_n\left(P_{W_n|S_n}\right)}{n} + \sqrt{\frac{\sigma^2 \log(1/\delta)}{2n}}$$

$$=$$

Generalized PAC-Bayes

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)|S_n]| \le \frac{h\left(P_{W_n|S_n}\right) - h(P_1)}{\eta n} + \frac{\eta B^2}{2\alpha} + \sqrt{\frac{\sigma^2 \log(1/\delta)}{2n}}$$

Basic examples

Relative entropy

$$\mathbb{E}[\mathrm{gen}(W_n,S_n)|S_n] \leq \sqrt{\frac{4\mathcal{D}_{\mathrm{KL}}\left(P_{W_n|S_n}\middle|P_0\right)\max_t\|c_t\|_\infty^2}{n}} + \sqrt{\frac{\sigma^2\log(\log n/\delta)}{2n}}$$

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

$$\mathbb{E}[\mathrm{gen}(W_n,S_n)|S_n] \leq \sqrt{\frac{4\left\|P_{W_n|S_n}-P_0\right\|_p^2\max_t\|c_t\|_q^2}{(p-1)n}} + \sqrt{\frac{\sigma^2\log(\log n/\delta)}{2n}}$$

p-norm with $p > 2$

$$\mathbb{E}[\mathrm{gen}(W_n,S_n)|S_n] \leq \frac{2p\left\|P_{W_n|S_n}-P_0\right\|_p^p\max_t\|c_t\|_q^q}{(p-1)n^{1/p}} + \sqrt{\frac{\sigma^2\log(\log n/\delta)}{2n}}$$

Basic examples

Relative entropy

$$\mathbb{E}[\mathrm{gen}(W_n, S_n)|S_n] \leq \sqrt{\frac{4\mathcal{D}_{\mathrm{KL}}\left(P_{W_n|S_n}\middle|P_0\right)\max_t \|c_t\|_\infty^2}{n}} + \sqrt{\frac{\sigma^2 \log(\log n/\delta)}{2n}}$$

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

$$\mathbb{E}[\mathrm{gen}(W_n, S_n)|S_n] \leq \sqrt{\frac{4\left\|P_{W_n|S_n} - P_0\right\|_p^2 \max_t \|c_t\|_q^2}{(p-1)n}} + \sqrt{\frac{\sigma^2 \log(\log n/\delta)}{2n}}$$

These norms remain meaningful for unbounded/heavy tailed losses

p-norm with $p > 2$

$$\mathbb{E}[\mathrm{gen}(W_n, S_n)|S_n] \leq \frac{2p\left\|P_{W_n|S_n} - P_0\right\|_p^p \max_t \|c_t\|_q^q}{(p-1)n^{1/p}} + \sqrt{\frac{\sigma^2 \log(\log n/\delta)}{2n}}$$

Basic examples

Relative entropy

$$\mathbb{E}[\mathrm{gen}(W_n,S_n)|S_n] \le \sqrt{\frac{4\mathcal{D}_{\mathrm{KL}}\left(P_{W_n|S_n}|P_0\right)\max_t\|c_t\|_\infty^2}{n}} + \sqrt{\frac{\sigma^2\log(\log n\,/\delta)}{2n}}$$

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

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

$$\mathbb{E}[\mathrm{gen}(W_n,S_n)|S_n] \le \sqrt{\frac{4\left\|P_{W_n|S_n}-P_0\right\|_p^2\max_t\|c_t\|_q^2}{(p-1)n}} + \sqrt{\frac{\sigma^2\log(\log n\,/\delta)}{2n}}$$

These norms remain meaningful for unbounded/heavy tailed losses

p-norm with $p > 2$

$$\mathbb{E}[\mathrm{gen}(W_n,S_n)|S_n] \le \frac{2p\left\|P_{W_n|S_n}-P_0\right\|_p^p\max_t\|c_t\|_q^q}{(p-1)n^{1/p}} + \sqrt{\frac{\sigma^2\log(\log n\,/\delta)}{2n}}$$

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

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

Theorem

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

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)]S_n| \le \sqrt{\frac{\frac{1}{\sigma^2}\mathbb{W}_2^2\left(P_{W_n|S_n}, P_0\right)\frac{1}{n}\sum_{t=1}^n \|c_t\|_{\sigma,*}^2}{n}} + \sqrt{\frac{\sigma^2\log(1/\delta)}{2n}}$$

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 $\mathcal{A}$,

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)]S_n| \leq \sqrt{\frac{\frac{1}{\sigma^{2}}\mathbb{W}_{2}^{2}\left(P_{W_n|S_n}, P_0\right)\frac{1}{n}\sum_{t=1}^{n}\|c_t\|_{\sigma,*}^{2}}{n}} + \sqrt{\frac{\sigma^{2}\log(1/\delta)}{2n}}$$

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 $\left|D^k f(w)\right| \leq \beta_k$.

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

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,

$$|\mathbb{E}[\mathrm{gen}(W_n, S_n)|S_n]| \leq \sqrt{\frac{8\beta^2 d \mathbb{W}_2^2\left(P_{W_n|S_n}, P_0\right)}{n}} + \sqrt{\frac{\sigma^2 \log(1/\delta)}{2n}}$$

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

Lemma

Suppose th for all k, all of

its partia $(w)| \le \beta_k$.

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,

$$|\mathbb{E}[\mathrm{gen}(W_n,S_n)|S_n]| \le \sqrt{\frac{8\beta^2 d\mathbb{W}_2^2\left(P_{W_n|S_n},P_0\right)}{n}} + \sqrt{\frac{\sigma^2\log(1/\delta)}{2n}}$$

What did we learn & what next?

- We can go beyond standard “information-theoretic” techniques!
- New since the COLT 2022 paper:
  - we can go beyond FTRL!
  - we can get high-probability bounds!
  - we can get data-dependent and parameter-free bounds!

What did we learn & what next?

- We can go beyond standard “information-theoretic” techniques!
- New since the COLT 2022 paper:
  - we can go beyond FTRL!
  - we can get high-probability bounds!
  - we can get data-dependent and parameter-free bounds!
- Many new possibilities:
  - data-dependent bounds? (non-trivial with current theory)
  - comparator-dependent bounds?
  - no need to worry about adaptivity!
  - no need to worry about implementability!

Thanks!!

Appendix

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 $\mathcal{D}_{\mathrm{KL}}$