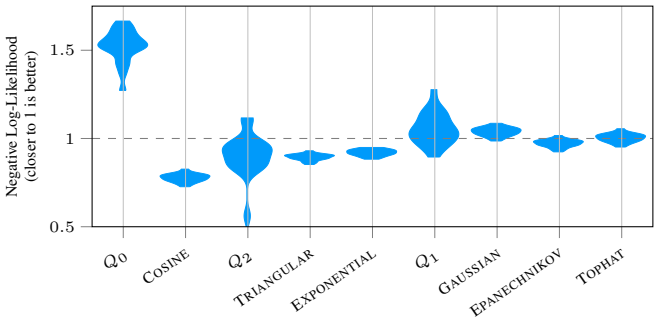
Supplementary Material for Boosted Density Estimation Remastered

Condition	Mean NLL ±95% CI
Q_0	1.5131 ±0.0459
COSINE	0.7734 ±0.0112
Q_2	0.8685 ±0.0946
TRIANGULAR	0.8898 ±0.0089
EXPONENTIAL	0.9154 ±0.0088
Q_1	1.0492 ±0.0437
GAUSSIAN	1.0333 ±0.0118
EPANECHNIKOV	0.9675 ±0.0105
TOPHAT	0.9983 ±0.0117

(a)



(b)

Figure 11: KDE comparison results. The conditions are in decreasing order with respect to the absolute difference of mean NLL and 1.

A. Epilogue

Results in the area of iterative approaches to density estimation can be analysed along three dimensions: how the *convergence* is characterised, the assumptions *assumptions* Unnormalised, and whether it is of direct relevance to current *empirical settings* for machine learning.

Regarding convergence, there are typically three degrees of formal results that are traditionally proven. The first is convergence without rates (Grover & Ermon, 2018; Dudík et al., 2004), the second are rates that are negligible with respect to recent results (including ours) (Rosset & Segal, 2002). The third and strongest results are explicit convergence rates. Some of the related approaches have an intractable objective and rather optimise a tractable surrogate bound. This is the case for variational inference, where the surrogate is the evidence lower-bound (Guo et al., 2016; Khan et al., 2016; Locatello et al., 2017; Miller et al., 2017). Because of the explicit gap to the intractable optimum, we do not mean to compare such approaches to ours, but most of the formal results in those papers yield sublinear convergence convergence, that is, of the form $I(P, Q_T) \in \mathrm{O}(\frac{1}{T})$ for some divergence measure I. In the rest of the related approaches, it quite remarkable that all of them use the same Frank–Wolfe-type update (1) (Li & Barron, 2000; Naito & Eguchi, 2013; Tolstikhin et al., 2017; Zhang, 2003). Until recently (Tolstikhin et al., 2017) all these other approaches essentially generated sublinear convergence rates (Li & Barron, 2000; Naito & Eguchi, 2013; Zhang, 2003). These can be compared to our rates from Theorem 15 and Theorem 18. We compare favourably with them from three standpoints: Firstly, all these algorithms

integrate calls to an oracle/subroutine that needs to solve a nested optimisation exactly — the constraint put on our oracle, the weak learner, appears much weaker. Secondly, all these algorithms integrate parameters whose computation would require the full knowledge of distributions (Naito & Eguchi, 2013; Zhang, 2003) or their parameterised space (Li & Barron, 2000). It is unclear how replacing these exact procedures by an approximation would impact convergence (Miller et al., 2017). In our case, Theorem 18 just operates on estimated parameters, straightforward to compute. Finally, previous works make more stringent structural assumptions restricting the form of the optimum (Li & Barron, 2000; Naito & Eguchi, 2013; Zhang, 2003), while we just assume that $c^\star$ is bounded, which puts a constraint — easily enforceable — on the proposals of the weak learner and not on the optimum.

On the topic of assumptions, the few previous approaches that manage to beat sublinear convergence to reach geometric convergence require very strong assumptions, such as the constraint that iterates are close enough to the optimum (Tolstikhin et al., 2017, Cor. 1, 2). In fact, in this latter work, the parameterisation of the weight α in (1) chosen for their experiments *implicitly imposes* the convergence of iterates to this optimum (Tolstikhin et al., 2017, §4). In our case, we have shown that equivalent convergence rates can be obtained without boosting (Corollary 6) but with an assumption which is used in (Tolstikhin et al., 2017, Cor. 1, Eq. 10), and is thus very strong. While this is not our main result, Corollary 6 is new and interesting in the light of Tolstikhin et al.'s results because (i) it does not make use of their convex mixture model and (ii) we do not have the additional technical requirement that $P(\mathrm{d}Q_{t-1}/\,\mathrm{d}P = 0) < \alpha_t$, that is, that roughly the mass where $\mathrm{d}Q_{t-1} = 0$ is bounded by the *leveraging* coefficient. We show that geometric convergence within reach with a much weaker assumptions than (Tolstikhin et al., 2017, Cor. 1, Eq. 10), in fact as weak as the weak learning assumption. To get our result, we need an additional assumption on the lower-boundedness of the log-errors ε_t almost everywhere via WDA. However, this is still not onerous given that we fit an exponential family, and in many interesting applications like image processing, $\mathcal{X}$ is closed so unless P is allowed to peak arbitrarily, we essentially get WDA for a reasonable Γ_ε.

Now, why is the assessment of all assumptions important in the light of experimental settings? Because it brings them to a trial by fire, as to whether results survive to experimental machine learning, with available information which is in general a partial estimated snapshot of the theory. It should be clear at this point that with the *sole* exception of a *subset of* variational approaches — which, again, settle for an explicitly tractable surrogate of the objective — *all previous approaches* would fail at this test, (Grover & Ermon, 2018; Guo et al., 2016; Li & Barron, 2000; Locatello et al., 2017; Naito & Eguchi, 2013; Tolstikhin et al., 2017; Zhang, 2003). They fail essentially because in practice we obviously would not have access to P to test assumptions nor carry out fine-grained optimisation. To our knowledge, our result in §C is the first attempt to provide an algorithm fully executable in current experimental learning settings and whose convergence relies on assumptions that would also easily be testable or enforceable empirically. Other approaches (variational inference and GANs) yield a black box sampler, which may be hard to train but are however fast to sample from in high dimensions (Guo et al., 2016; Khan et al., 2016; Locatello et al., 2017; Miller et al., 2017; Tolstikhin et al., 2017). This is clearly where bottleneck of our

theory lies. A solution to the sampling problem is therefore all that is conceivably preventing our approach from application to similar, high dimensional settings.

B. The error term

Recall the reparameterised variational problem from §2

$$\underset{u}{\text{minimise}} \quad J(u) \stackrel{\text{def}}{=} \mathrm{E}_Q\, f^* \circ f' \circ u - \mathrm{E}_P\, f' \circ u \quad \text{subject to} \quad u \in \mathcal{F}. \tag{V}$$

The solution to (V) easily follows when $\mathcal{F}$ is a large enough set of measurable functions (Nowozin et al., 2016; Nguyen et al., 2010; Grover & Ermon, 2018). However when $\mathcal{F}$ is a more constrained class, a stronger result is necessary. Assume $\mathcal{F}$ is a subset of the normed space, $(\mathscr{F}, |\cdot|)$. Let $\mathscr{F}^*$ be its continuous dual. The Fréchet normal cone (also called prenormal cone) of $\mathcal{F} \subseteq \mathscr{F}$ at $u \in \mathscr{F}$ is

$$\mathrm{N}_{\mathcal{F}}(u) \stackrel{\text{def}}{=} \left\{ u^* \in \mathscr{F}^* : \limsup_{\mathcal{F} \supseteq (v) \to u} \frac{\langle u^*, v - u \rangle}{|v - u|} \leq 0 \right\}.$$

When $\mathcal{F}$ is convex, $\mathrm{N}_{\mathcal{F}}(u)$ is the ordinary normal cone.

Theorem 1. *Assume $f : \mathbb{R}_+ \to \mathbb{R}_+$ is strictly convex and twice differentiable, and $\mathscr{F}$ is a normed space of functions $\mathcal{X} \to \operatorname{int}(\operatorname{dom} f)$. Let $\mathcal{F} \subseteq \mathscr{F}$ and $\bar{u} \in \arg\min_{u \in \mathcal{F}} J(u)$. If J is finite on a neighbourhood of $\bar{u}$, then*

$$\bar{u} \in \frac{\mathrm{d}P}{\mathrm{d}Q} - \mathrm{N}_{\mathcal{F}}(\bar{u}).$$

If, in addition, $\mathcal{F}$ is convex with $\mathrm{d}P/\mathrm{d}Q \in \operatorname{int} \mathcal{F}$, then $\bar{u} = \mathrm{d}P/\mathrm{d}Q$.

Proof. Because f is twice differentiable on $\operatorname{int}(\operatorname{dom} f)$, and J is finite on a neighbourhood of $\bar{u}$, J is Fréchet differentiable at $\bar{u}$ with

$$\begin{aligned} J'(\bar{u}) &= ((f^*)' \circ f' \circ \bar{u}) \cdot (f'' \circ \bar{u}) \cdot \mathrm{d}Q - (f'' \circ \bar{u}) \cdot \mathrm{d}P. \\ &= \bar{u} \cdot (f'' \circ \bar{u}) \cdot \mathrm{d}Q - (f'' \circ \bar{u}) \cdot \mathrm{d}P, \end{aligned}$$

where $(f^*)' = (f')^{-1}$ since f is strictly convex. By hypothesis J attains its minimum on $\mathcal{F}$ at $\bar{u}$, thus Fermat's rule (Penot, 2012, Thm. 2.97, p. 170) yields

$$\begin{aligned} 0 \in J'(\bar{u}) + \mathrm{N}_{\mathcal{F}}(\bar{u}) &\iff 0 \in \bar{u} \cdot (f'' \circ \bar{u}) \cdot \mathrm{d}Q - (f'' \circ \bar{u}) \cdot \mathrm{d}P + \mathrm{N}_{\mathcal{F}}(\bar{u}) \\ &\iff 0 \in \bar{u} - \frac{\mathrm{d}P}{\mathrm{d}Q} + \frac{1}{(f'' \circ \bar{u}) \cdot \mathrm{d}Q} \cdot \mathrm{N}_{\mathcal{F}}(\bar{u}) \\ &\iff \bar{u} \in \frac{\mathrm{d}P}{\mathrm{d}Q} - \mathrm{N}_{\mathcal{F}}(\bar{u}), \end{aligned}$$

where the final biconditional follows since $\mathrm{N}_{\mathcal{F}}(\bar{u})$ is a cone.

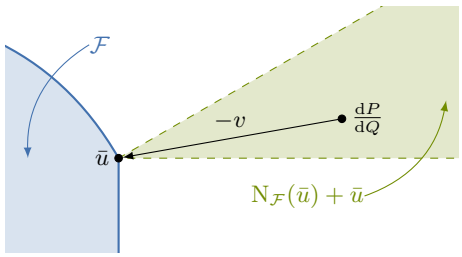


Figure 12: Illustration of Theorem 1 wherein there exists $v \in \mathrm{N}_{\mathcal{F}}(\bar{u})$ which pulls the unconstrained minimiser, $\mathrm{d}P/\mathrm{d}Q$, onto the constrained minimiser, $\bar{u}$.

Now, suppose $\mathrm{d}P/\mathrm{d}Q \in \operatorname{int} \mathcal{F}$ with $\mathcal{F}$ convex. Then the Fréchet cone becomes usual normal cone (Penot, 2012, Ex. 6, p. 174),

$$\mathrm{N}_{\mathcal{F}}(\bar{u}) \stackrel{\text{def}}{=} \{u^* \in \mathcal{F}^* : \forall v \in \mathcal{F} : \ \langle u^*, v - \bar{u} \rangle \leq 0\}.$$

It's immediate from the definition that $\mathrm{N}_{\mathcal{F}}$ always contains 0. We use a contradiction to show that $\mathrm{N}_{\mathcal{F}}(\bar{u}) \subseteq \{0\}$. Take $z^* \neq 0 \in \mathrm{N}_{\mathcal{F}}(\bar{u})$. Let $\mathcal{F}_{\bar{u}} \stackrel{\text{def}}{=} \mathcal{F} - \bar{u}$. First note that $\bar{u} \in \operatorname{int} \mathcal{F}$ implies $0 \in \operatorname{int} \mathcal{F}_{\bar{u}}$. Thus there is a closed symmetric neighbourhood U with $0 \in U \subseteq \operatorname{int} \mathcal{F}_{\bar{u}}$. The Hahn–Banach strong separation theorem (Penot, 2012, Thm. 1.79, p. 55) guarantees the existence of a vector $u \in U$ such that

$$\langle z^*, u \rangle > 0 \iff \exists v \in \mathcal{F} : \ \langle z^*, v - \bar{u} \rangle > 0,$$

contradicting the assumption $z^* \in \mathrm{N}_{\mathcal{F}}$. Thus $\mathrm{N}_{\mathcal{F}}(\mathrm{d}P/\mathrm{d}Q) = \{0\}$. ■

The set $\mathrm{N}_{\mathcal{F}}(\bar{u})$ can be thought of as containing a direction v that pulls $\mathrm{d}P/\mathrm{d}Q$ to the constrained minimiser $\bar{u}$. This is illustrated in Figure 12.

Theorem 1 also gives us give a more explicit characterisation of the error term in §3 since

$$\exists v_t \in \mathrm{N}_{\mathcal{F}}(d_t) : \ d_t = \frac{\mathrm{d}P}{\mathrm{d}Q_t} \cdot \varepsilon_t = \frac{\mathrm{d}P}{\mathrm{d}Q_t} - v_t, \iff \varepsilon_t = 1 - \frac{\mathrm{d}Q_{t-1}}{\mathrm{d}P} \cdot v_t.$$

C. Boosting with estimates

In practice, we do not have access to P and we rather sample from Q. We thus assume the possibility to sample[8] P and $Q_\cdot$ to compute all needed estimates of μ_P and $\nu_{Q_\cdot}$. So let us assume that the weak learner has access to a sampler of P and a sampler of $Q_\cdot$, "SAMPL". SAMPL takes as input a distribution and a natural m; it samples from the distribution and returns an i.i.d. sample of size m. It does so separately for P and Q, with separate sizes m_P and m_Q for the respective samples. The full Figure 2 is very similar to Figure 1 if we except

[8] We could also assume the availability of training samples, in particular for P as is usually carried out.

the fact that the weak learner also returns an estimate for $\nu_{Q_{t-1}}$. To analyze Figure 2 requires however more than just the boosting material developed so far, since nothing guarantees that estimates of μ_P and $\nu_{Q_\cdot}$ meeting WLA would imply μ_P and $\nu_{Q_\cdot}$ meeting WLA as well. We therefore replace WLA by one which relies on *estimates* computed over large enough samples. We call it the *Empirical* Weak Learner Assumption (EWLA). It involves two additional parameters, $\kappa^\star = \nu_{c^\star}(1-\nu_{c^\star})/2\ (> 0)$, where $\nu_{c^\star}$, which depends only on $c^\star$, is defined in (43), and some $0 < \delta \leq 1$.

Assumption 3 (Empirical Weak Learner Assumption)**.** *There exists $\gamma_P, \gamma_Q \in (0,1]$ such that the following holds: at each iteration $t = 1, 2, ..., T$,* WEAKLEARN *estimates $\hat{\mu}_P$ and $\hat{\nu}_{Q_{t-1}}$ respectively from*

$$\text{SAMPL}(P, m_P) \quad \textit{and} \quad \text{SAMPL}(Q_{t-1}, m_Q)$$

with m_P, m_Q satisfying

$$m_P \geq \frac{1}{(\kappa^\star \gamma_P)^2} \log \frac{4T}{\delta} \quad \textit{and} \quad m_Q \geq \frac{1}{(\kappa^\star \gamma_Q)^2} \log \frac{4T}{\delta}, \qquad (\text{EWLA}_{\delta,T})$$

and returns, along with $\hat{\nu}_{Q_{t-1}}$, c_t satisfying $\hat{\mu}_P \geq \gamma_P$ and $\hat{\nu}_{Q_{t-1}} \geq \gamma_Q$.

The weak learner thus also take as input δ and T, as displayed in Figure 2. We emphasize the fact that $\text{EWLA}_{\delta,T}$ is just assuming the ability for WEAKLEARN to have i.i.d. samples from P and Q and get a classifier c_t that *empirically* satisfies WLA. Since we still focus on the decrease of $\text{KL}(P, Q_\cdot)$, one might expect this to weaken our results, which is indeed the case, *but* we can show that *only* constants are *slightly* affected, thereby not changing significantly convergence rates. We provide in one theorem the reframing of both Theorem 15 and Theorem 17. In the same way as we did for Theorem 17, whenever we are in the clamped regime for α_t, we let $\hat{\delta}_{t-1} \geq 0$ be defined from $\hat{\nu}_{Q_{t-1}} = (1 + \hat{\delta}_{t-1})\nu_{c^\star}$.

Theorem 23. *Suppose $\text{EWLA}_{\delta,T}$ holds. Then with probability of at least $1-\delta$,*

$$\forall t = 1, 2, ..., T: \ \text{KL}(P, Q_t) \leq \text{KL}(P, Q_{t-1}) - \Delta_t,$$

where

$$\Delta_t \stackrel{\text{def}}{=} \begin{cases} \frac{\hat{\mu}_P}{16} \log\left(\frac{1+\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right) & \textit{in the non-clamped regime,} \\ \frac{\hat{\mu}_P c^\star}{2} + \nu_{c^\star}^2 \cdot \left(\frac{1}{4} + \frac{\hat{\delta}_{t-1}}{1-\nu_{c^\star}^2}\right) & \textit{otherwise.} \end{cases}$$

D. Proofs of formal results

Proposition 2. *The normalisation factors can be written recursively with $Z_t = Z_{t-1} \cdot \text{E}_{Q_{t-1}} d_t^{\alpha_t}$.*

Proof. We just need to write

$$\frac{Z_t}{Z_{t-1}} = \frac{1}{Z_{t-1}} \int \text{d}\tilde{Q}_t = \frac{1}{Z_{t-1}} \int d_t^{\alpha_t} \, \text{d}\tilde{Q}_{t-1} = \int d_t^{\alpha_t} \, \text{d}Q_{t-1} = \text{E}_{Q_{t-1}} d_t^{\alpha_t} \qquad (10)$$

thus $Z_t = Z_{t-1} \cdot \text{E}_{Q_{t-1}} d_t^{\alpha_t}$. ■

Algorithm 1 AdaBoDE

Input: distributions P, Q_0, natural T;
for $t = 1$ **to** T **do**
  $c_t \leftarrow \textsc{WeakLearn}(P, Q_{t-1})$;
  Pick $\nu_{Q_{t-1}}$ as in (8) and α_t as
  $\alpha_t \leftarrow \min\left\{1, \frac{1}{2c^\star} \log\left(\frac{1+\nu_{Q_{t-1}}}{1-\nu_{Q_{t-1}}}\right)\right\}$;
  Pick Q_t as in (4) with $d_t \stackrel{\text{def}}{=} \exp \circ c_t$;
end for
Return: Q_T

Algorithm 2 AdaBoDE.EST

Input: distributions P, Q_0, $T \in \mathbb{N}_*$, $0 < \delta \leq 1$;
for $t = 1$ **to** T **do**
  $(c_t, \hat{\nu}_{Q_{t-1}}) \leftarrow \textsc{WeakLearn}(P, Q_{t-1}, \delta, T)$;
  $\alpha_t \leftarrow \min\left\{1, \frac{1}{2c^\star} \log\left(\frac{1+\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right)\right\}$;
  Pick Q_t as in (4) with $d_t \stackrel{\text{def}}{=} \exp(c_t)$;
end for
Return: Q_T

Proposition 3. *Let Q_t be defined via (4) with a sequence of binary classifiers $c_1, \ldots, c_t \in \mathcal{C}(\mathcal{X})$. Then Q_t is an exponential family distribution with natural parameter $\alpha \stackrel{\text{def}}{=} (\alpha_1, \ldots, \alpha_t)$ and sufficient statistic $c(x) \stackrel{\text{def}}{=} (c_1(x), \ldots, c_t(x))$.*

Proof. We can convert the binary classifiers $c_1, \ldots, c_t \in \mathcal{C}(\mathcal{X})$ to a sequence of density ratios (d_i) using the connections in §2, which yields

$$d_i^{\alpha_i} \stackrel{\text{def}}{=} (\varphi \circ \sigma \circ c_i)^{\alpha_i} = \exp \circ (\alpha_i c_i).$$

In this setting, the multiplicative density at round t is

$$\begin{aligned} \mathrm{d}Q_t(x) &\stackrel{(4)}{=} \frac{1}{\int \prod_{i=1}^t d_i^{\alpha_i} \, \mathrm{d}Q_0} \prod_{i=1}^t d_i^{\alpha_i} \, \mathrm{d}Q_i(x) \\ &= \exp\left(\sum_{i=1}^t \alpha_i c_i(x) - C(\alpha)\right) \mathrm{d}Q_0(x), \end{aligned}$$

with $\alpha \stackrel{\text{def}}{=} (\alpha_1, \ldots, \alpha_t)$ and $C(\alpha) = \log \int \exp\left(\sum_{i=1}^t \alpha_i c_i\right) \mathrm{d}Q_0$, which is an exponential family distribution with natural parameter α, sufficient statistic $c(x) \stackrel{\text{def}}{=} (c_1(x), \ldots, c_t(x))$, cumulant function $C(\alpha)$, reference measure Q_0. We note that in the general case, it may be the case that for some non-all-zero constants $z_0, z_1, \ldots, z_t \in \mathbb{R}$, we have $z_0 = \sum_{i=1}^t z_i c_i(x)$, that is, the representation is not minimal. ■

Lemma 4. *For any $\alpha_t \in [0, 1]$ and $\varepsilon_t \in [0, +\infty)^{\mathcal{X}}$ we have:*

$$\exp\Big(\alpha_t\big(\mathrm{E}_{Q_{t-1}} \log \varepsilon_t - \mathrm{rKL}(P, Q_{t-1})\big)\Big) \leq \frac{Z_t}{Z_{t-1}} \leq \left(\mathrm{E}_P \, \varepsilon_t\right)^{\alpha_t}.$$

Proof. Since $\alpha_t \in [0, 1]$, by Jensen's inequality it follows that

$$\mathrm{E}_{Q_{t-1}} \, d_t^{\alpha_t} \leq \left(\mathrm{E}_{Q_{t-1}} \, d_t\right)^{\alpha_t} = \left(\int \frac{\mathrm{d}P}{\mathrm{d}Q_{t-1}} \cdot \varepsilon_t \, \mathrm{d}Q_{t-1}\right)^{\alpha_t} = \left(\mathrm{E}_P \, \varepsilon_t\right)^{\alpha_t}. \tag{11}$$

The upper bound on Z_t/Z_{t-1} follows:

$$\frac{Z_t}{Z_{t-1}} \overset{(10)}{=} \mathrm{E}_{Q_{t-1}}\, d_t^{\alpha_t} \overset{(11)}{\leq} (\mathrm{E}_P\, \varepsilon_t)^{\alpha_t}.$$

For the lower bound on Z_t/Z_{t-1}, note that

$$\begin{aligned}\log\left(\frac{Z_t}{Z_{t-1}}\right) &\overset{(10)}{=} \log \mathrm{E}_{Q_{t-1}}\, d_t^{\alpha} \\ &\geq \alpha_t\, \mathrm{E}_{Q_{t-1}} \log d_t \\ &= \alpha_t\, \mathrm{E}_{Q_{t-1}} \left[\log \varepsilon_t + \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q_{t-1}}\right)\right],\end{aligned}$$

which implies the lemma. ■

The error term allows us to bound the KL divergence of P from Q_t as follows.

Lemma 5. *For any $\alpha_t \in [0,1]$, letting Q_t, Q_{t-1} as in (3), we have:*

$$\forall d_t \in \mathcal{R}(\mathcal{X}):\ \mathrm{KL}(P, Q_t|_{\alpha_t}) \leq (1-\alpha_t)\,\mathrm{KL}(P, Q_{t-1}) + \alpha_t(\log \mathrm{E}_P\, \varepsilon_t - \mathrm{E}_P \log \varepsilon_t). \tag{12}$$

where $d_t = \mathrm{d}P/\mathrm{d}Q_{t-1} \cdot \varepsilon_t$.

Proof. First note that

$$\mathrm{d}Q_t = \frac{1}{Z_t}\,\mathrm{d}\tilde{Q}_t = \frac{1}{Z_t} d_t^{\alpha_t}\,\mathrm{d}\tilde{Q}_{t-1} = \frac{Z_{t-1}}{Z_t} d_t^{\alpha_t}\,\mathrm{d}Q_{t-1}. \tag{13}$$

Now consider the following two identities:

$$-\alpha_t \log \mathrm{E}_P\, \varepsilon_t \leq \log\left(\frac{Z_{t-1}}{Z_t}\right), \tag{14}$$

which follows from Lemma 4, and

$$\begin{aligned}&\int \left(\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q_{t-1}}\right) - \alpha_t \log d_t\right) \mathrm{d}P \\ &\quad= \int \left(\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q_{t-1}}\right) - \alpha_t \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q_{t-1}}\right) - \alpha_t \log \varepsilon_t\right) \mathrm{d}P \\ &\quad= (1-\alpha_t) \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q_{t-1}}\right) \mathrm{d}P - \alpha_t \int \log \varepsilon_t\, \mathrm{d}P \\ &\quad= (1-\alpha_t)\,\mathrm{KL}(P, Q_{t-1}) - \alpha_t\, \mathrm{E}_P \log \varepsilon_t.\end{aligned} \tag{15}$$

Then

$$\begin{aligned}
\mathrm{KL}(P, Q_t) &= \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q_t}\right) \mathrm{d}P \\
&\overset{(13)}{=} \int \left(\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q_{t-1}}\right) - \log\left(\frac{Z_{t-1}}{Z_t} d_t^{\alpha_t}\right)\right) \mathrm{d}P \\
&= \underbrace{\int \left(\log\left(\frac{\mathrm{d}P}{\mathrm{d}Q_{t-1}}\right) - \alpha_t \log d_t\right) \mathrm{d}P}_{(15)} - \underbrace{\log\left(\frac{Z_{t-1}}{Z_t}\right)}_{(14)} \\
&\leq (1 - \alpha_t)\, \mathrm{KL}(P, Q_{t-1}) + \alpha_t (\log \mathrm{E}_P\, \varepsilon_t - \mathrm{E}_P \log \varepsilon_t),
\end{aligned}$$

as claimed. ■

Corollary 6. *For any $\alpha_t \in [0, 1]$ and $\varepsilon_t \in [0, +\infty)^{\mathcal{X}}$, letting Q_t as in (4) and R_t from (6). If R_t satisfies*

$$\mathrm{KL}(P, R_t) \leq \gamma\, \mathrm{KL}(P, Q_{t-1})$$

for $\gamma \in [0, 1]$, then

$$\mathrm{KL}(P, Q_t|_{\alpha_t}) \leq (1 - \alpha_t(1 - \gamma))\, \mathrm{KL}(P, Q_{t-1}). \tag{16}$$

Proof. We first show

$$\mathrm{KL}(P, Q_t|_{\alpha_t}) \leq (1 - \alpha_t)\, \mathrm{KL}(P, Q_{t-1}) + \alpha_t\, \mathrm{KL}(P, R_t). \tag{17}$$

By definition $\varepsilon_t = \mathrm{d}R_t / \mathrm{d}P$. The rightmost term in (12) reduces as follows

$$\begin{aligned}
\log \mathrm{E}_P\, \varepsilon_t - \mathrm{E}_P \log \varepsilon_t &= \log \int \frac{\mathrm{d}\tilde{R}_t}{\mathrm{d}P}\, \mathrm{d}P - \int \log\left(\frac{\mathrm{d}\tilde{R}_t}{\mathrm{d}P}\right) \mathrm{d}P \\
&= \log \int \mathrm{d}\tilde{R}_t + \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}\tilde{R}_t}\right) \mathrm{d}P \\
&= \int \left(\log\left(\frac{\mathrm{d}P}{\mathrm{d}\tilde{R}_t}\right) + \log \int \mathrm{d}\tilde{R}_t\right) \mathrm{d}P \\
&= \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}\tilde{R}_t} \cdot \int \mathrm{d}\tilde{R}_t\right) \mathrm{d}P \\
&= \int \log\left(\frac{\mathrm{d}P}{\frac{1}{\int \mathrm{d}\tilde{R}_t}\, \mathrm{d}\tilde{R}_t}\right) \mathrm{d}P,
\end{aligned}$$

which completes the proof of (17). The proof of (16) is then immediate. ■

Define WEAKLEARN the weak learner which, taking P and Q_{t-1} as input, delivers c_t satisfying the conditions of WLA. In the boosting theory, which involves a supervised

learning problem, there is one condition instead of two as in WLA: given a distribution D over $\mathcal{X} \times \{-1, +1\}$, we rather require from the weak learner, WEAKLEARN*, that

$$\exists \gamma \in (0, 1] : \ \frac{1}{c^\star} \mathrm{E}_D \, y \cdot c_t \geq \gamma,$$

where y denotes the class, mapping $\mathcal{X} \to \{-1, +1\}$. While is seems rather intuitive that we can craft WEAKLEARN* from WEAKLEARN, it is perhaps less intuitive as to whether the same can be done for the reverse direction. We now show that it is indeed the case and WLA and WLA* are in fact equivalent.

Lemma 7. *Suppose $\gamma_P = \gamma_Q = \gamma$ in WLA and WLA*, without loss of generality. Then there exists* WEAKLEARN *satisfying WLA iff there exists* WEAKLEARN* *satisfying WLA*.*

Proof. To simplify notations, we suppose without loss of generality that $\mathcal{C}(\mathcal{X}) \subseteq [-1, 1]^{\mathcal{X}}$.

($\Rightarrow$) Let D be a distribution on $\mathcal{X} \times \{-1, +1\}$. It can be factored as a triple (π, P, N) where P is a distribution over the positive examples, N is a distribution over negative examples and π is the mixing probability, $\pi \stackrel{\text{def}}{=} \Pr_D[y = +1]$. Now, feed P and N in lieu of P and Q_t, repectively. We get c_t which, from WLA, satisfies $\mathrm{E}_N[-c_t] \geq \gamma$ and $\mathrm{E}_P[c_t] \geq \gamma$, which implies

$$\mathrm{E}_D[yc_t] = \pi \, \mathrm{E}_P[c_t] + (1 - \pi) \, \mathrm{E}_N[-c_t] \geq \pi\gamma + (1 - \pi)\gamma = \gamma$$

and we get our weak learner WEAKLEARN* satisfying WLA*.

($\Leftarrow$) We create a two-class classification problem in which observations from P have positive class $y = +1$, observations from Q_t have negative class $y = -1$ and there is a special observation $x^* \in \mathcal{X}$ which is equally present with probability $1 - 2\pi$ in both the positive and negative class. Hence, we are artificially increasing the difficulty of the problem by making its Bayes optimum worse. Obviously, WLA* having to hold under any distribution, it will have to hold under the distribution D that we create. To explicit D, consider $\pi \in [0, 1/2]$ and the following sampler for D:

- sample $z \in [0, 1]$ uniformly;
  - if $z \leq (1 - 2\pi)/2$ return $(x^*, +1)$;
  - else if $z \leq 1 - 2\pi$ return $(x^*, -1)$;
  - else if $z \leq 1 - \pi$, return $(x \sim P, +1)$;
  - else return $(x \sim Q_t, -1)$;

Let D denote the distribution induced on $\mathcal{X} \times \{-1, +1\}$. Remark that the error of Bayes optimum is at least $1/2 - \pi$. Let c_t returned by WEAKLEARN*. We have because of WLA*,

$$\mathrm{E}_D[yc_t] = \pi(\mu_P + \nu_{Q_{t-1}}) + \left(\frac{1 - 2\pi}{2} - \frac{1 - 2\pi}{2}\right) c_t(x^*) = \pi(\mu_P + \nu_{Q_{t-1}}) \geq \gamma$$

Consider

$$\pi = \frac{\gamma}{1+\gamma}$$

Which makes

$$\mu_P + \nu_{Q_{t-1}} \geq 1 + \gamma.$$

It easily comes that if $\mu_P < \gamma$, then we must have $\nu_{Q_{t-1}} > 1$, which is not possible, and similarly if $\nu_{Q_{t-1}} < \gamma$, then we must have $\mu_P > 1$, which is also impossible. Therefore we have both $\mu_P \geq \gamma$ and $\nu_{Q_{t-1}} \geq \gamma$, and we get our weak learner WEAKLEARN meeting WLA, as claimed. ■

D.0.1. PROOF OF THEOREM 15

The proof of Theorem 15 is achieved in two steps: (i) any c_t meeting WLA can be transformed through scaling into a classifier that we call *Properly Scaled* without changing it satisfying WLA (for the same parameters γ_P, γ_Q), (ii) Theorem 15 holds for such Properly Scaled classifiers.

Definition 4. *The classifier c_t is said to be Properly Scaled (PS) if it meets:*

$$\exp(2c^\star) \leq 2 + \mu_P c^\star \tag{PS.1}$$

$$\mathrm{E}_{Q_{t-1}} \exp(c_t) \leq \exp\left(\frac{\mu_P c^\star}{4}\right). \tag{PS.2}$$

Hence, we first show how any classifier meeting WLA can be made PS *without* changing μ_P nor $\nu_{Q_{t-1}}$ (hence, still meeting WLA), modulo a simple positive scaling. The proof involves a reverse of Jensen's inequality which is much simpler than previous bounds (Simić, 2009a;b) and of independent interest.

Our proof will equivalently give upper bounds on $c^\star$ that make c_t PS. We note that our proof is constructive, that is, we give eligible upper bounds for $c^\star$. The proof of Theorem 12 is split in several lemmata, the first of which is straightforward since $\mu_P \geq 0$ under WLA.

Lemma 8. *Suppose c_t meets WLA. Then, (PS.1) holds for any $c^\star \leq \log(2)/2$.*

To prove how to satisfy (PS.2), we use the notions of Bregman divergences and Bregman information. For $\varphi : \mathbb{R} \to \mathbb{R}$ convex differentiable with derivative φ', we define the Bregman divergence with generator φ as $D_\varphi(z \| z') = \varphi(z) - \varphi(z') - (z - z')\varphi'(z')$. Following (Banerjee et al., 2005), we define the *minimal Bregman information* of (c_t, Q_{t-1}) (or just *Bregman information* for short) relative to φ as

$$I_\varphi(c_t; Q_{t-1}) \stackrel{\text{def}}{=} \mathrm{E}_{Q_{t-1}}[D_\varphi(c_t \| \mathrm{E}_{Q_{t-1}} c_t)].$$

The Bregman information is a generalization of the variance for which $\varphi(z) = z^2$. Jensen's inequality would give us a lowerbound, but we need an *upperbound*. We devise for this objective a reverse of Jensen's inequality. We suppose that c_t takes values in $[a, b]$, where we would thus have $|a|$ or b which would be $c^\star$.

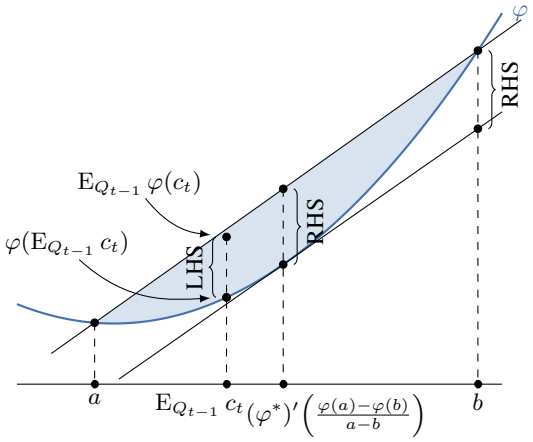


Figure 13: Proof of Lemma 9. The tangent represented is parallel to the chord linking $(a, \varphi(a))$ and $(b, \varphi(b))$. The LHS of (19) can be represented by the vertical difference labelled (always in the blue area). The RHS of (19), as a Bregman divergence, is the difference between the φ and its tangent at $(\varphi^*)'(\frac{\varphi(a)-\varphi(b)}{a-b})$, measured at either a, or b (pictured).

Lemma 9. *(Reverse of Jensen's inequality) Suppose φ strictly convex differentiable and $c_t(x) \in [a, b]$ for all $x \in \mathcal{X}$. Then,*

$$I_\varphi(c_t; Q_{t-1}) \leq D_\varphi\left(u \middle\| (\varphi^*)'\left(\frac{\varphi(a)-\varphi(b)}{a-b}\right)\right), \tag{19}$$

where u can be chosen to be a or b.

Proof. The proof is in fact straightforward, as illustrated in Figure 13. ■

We now show how to satisfy (PS.2).

Lemma 10. *Suppose c_t meets WLA and*

$$c^\star \leq \left(\frac{\gamma_P}{4} + \frac{1-\exp(-2\gamma_Q)}{2}\right).$$

Then

$$\mathrm{E}_{Q_{t-1}} \exp(c_t) \leq \exp\left(\frac{\mu_P c^\star}{4}\right),$$

that is, (PS.2) *holds.*

Proof. Consider $\varphi(z) = \exp(z)$ and so $\varphi^*(z) = z\log z - z$ in Lemma 9. Suppose without loss of generality that $a = -c^\star, b = c^\star$. We get

$$\begin{aligned} I_{\exp}(c_t; Q_{t-1}) &= \mathrm{E}_{Q_{t-1}} \exp(c_t) - \exp \mathrm{E}_{Q_{t-1}} c_t \\ &\leq D_\varphi\left(c^\star \middle\| (\varphi^*)'\left(\frac{\varphi(c^\star) - \varphi(-c^\star)}{2c^\star}\right)\right). \end{aligned}$$

Now, we just need to ensure that

$$D_\varphi\left(c^\star \middle\| (\varphi^*)'\left(\frac{\varphi(c^\star) - \varphi(-c^\star)}{2c^\star}\right)\right) \leq \exp\left(\frac{\mu_P c^\star}{4}\right) - \exp(-\nu_{Q_{t-1}} c^\star), \qquad (20)$$

as indeed we shall then have, because of WLA,

$$\begin{aligned} \mathrm{E}_{Q_{t-1}} \exp(c_t) &\leq \exp\left(\frac{\gamma_P c^\star}{4}\right) - \exp(-\gamma_Q c^\star) + \exp \mathrm{E}_{Q_{t-1}} c_t \\ &\leq \exp\left(\frac{\mu_P c^\star}{4}\right) - \exp(-\nu_{Q_{t-1}} c^\star) + \exp(-\nu_{Q_{t-1}} c^\star) \\ &= \exp\left(\frac{\mu_P c^\star}{4}\right), \end{aligned}$$

which is the statement of the lemma.

Proposition 11. *Pick* $\varphi \stackrel{\text{def}}{=} \exp$. *If* $|z| \leq 2$, *then*

$$D_\varphi\left(z \middle\| (\varphi^*)'\left(\frac{\varphi(z) - \varphi(-z)}{2z}\right)\right) \leq z^2.$$

Proof. Equivalently, we need to show

$$\begin{aligned} z^2 \geq \exp(z)\left(\frac{1}{2} - \frac{1}{2z}\right) + \exp(-z)\left(\frac{1}{2} + \frac{1}{2z}\right) \\ + \left(\frac{\exp(z) - \exp(-z)}{2z}\right) \log\left(\frac{\exp(z) - \exp(-z)}{2z}\right). \end{aligned}$$

We split the proof in two. First, let us fix

$$g_1(z) \stackrel{\text{def}}{=} \frac{2\big(\exp(z)\big(\frac{1}{2} - \frac{1}{2z}\big) + \exp(-z)\big(\frac{1}{2} + \frac{1}{2z}\big)\big)}{z^2}.$$

We remark that

$$g_1'(z) = \frac{\exp(-z)(-z^2 - 3z - 3 + \exp(2z)(z^2 - 3z + 3))}{2z^4}.$$

We then remark that, letting $g_2(z) \stackrel{\text{def}}{=} -z^2 - 3z - 3 + \exp(2z)(z^2 - 3z + 3)$,

$$\begin{aligned}
g_2(z) &= -z^2 - 3z - 3 + \sum_{k\geq 0} \frac{2^k z^{k+2}}{k!} - \sum_{k\geq 0} \frac{3 \cdot 2^k z^{k+1}}{k!} + \sum_{k\geq 0} \frac{3 \cdot 2^k z^k}{k!} \\
&= -z^2 - 3z - 3 + \sum_{k\geq 2} \frac{2^{k-2} z^k}{(k-2)!} - \sum_{k\geq 1} \frac{3 \cdot 2^{k-1} z^k}{(k-1)!} + \sum_{k\geq 0} \frac{3 \cdot 2^k z^k}{k!} \\
&= \sum_{k\geq 2} \left(\frac{2^{k-2}}{(k-2)!} - \frac{3 \cdot 2^{k-1}}{(k-1)!} + \frac{3 \cdot 2^k}{k!} \right) z^k \\
&= \sum_{k\geq 5} \left(\frac{2^{k-2}}{(k-2)!} - \frac{3 \cdot 2^{k-1}}{(k-1)!} + \frac{3 \cdot 2^k}{k!} \right) z^k \\
&= \sum_{k\geq 5} \big(k^2 - 7k + 12\big) \frac{2^{k-2} z^k}{k!}.
\end{aligned}$$

We then check that $z \mapsto z^2 - 7z + 12 < 0$ only for $z \in (3, 4)$. That is, it is never negative over naturals so $g_2(z) \geq 0, \forall z \geq 0$. We also check that $\lim_0 g_1'(z) = 0$ and so $g_1(z)$ is increasing for $z \geq 0$. Finally,

$$g_1(2) = \frac{1}{2} \cdot \left(\frac{\exp(2)}{4} + \frac{3}{4 \exp(2)} \right) < \frac{7.81}{8} < 1,$$

which shows that

$$\forall |z| \leq 2 : \ \exp(z) \left(\frac{1}{2} - \frac{1}{2z} \right) + \exp(-z) \left(\frac{1}{2} + \frac{1}{2z} \right) \leq \frac{z^2}{2}. \tag{21}$$

(The analysis for $z < 0$ uses the fact that the function is even.) We now show that we have

$$\forall z \in [-2, 2] : \ \frac{\exp(z) - \exp(-z)}{2z} \leq 1 + \frac{z^2}{4}. \tag{22}$$

Hence, we want to show that $\exp(z) \leq \exp(-z) + 2z + z^3/2$ for $z \in [-2, 2]$. We now have $\exp(-z) \geq 1 - z + z^2/2 - z^3/6 + z^4/24 - z^5/120$ for $z \geq 0$, so we just need to show $\exp(z) \leq 1 - z + z^2/2 - z^3/6 + z^4/24 - z^5/120 + 2z + z^3/2 = 1 + z + z^2/2 + z^3/3 + z^4/24 - z^5/120 + 2z + z^3/2$ for $z \in [0, 2]$ (we will then use the fact that both functions in (22) are even), which simplifies, using Taylor series for $\exp$, in showing

$$\forall z \in [0, 2] : \ \sum_{k\geq 6} \frac{z^k}{k!} \leq \frac{z^3}{6} - \frac{z^5}{60},$$

or after dividing both sides by $z^3 > 0$ (the inequality is obviously true for $z = 0$),

$$\forall z \in (0, 2] : \ \sum_{k\geq 6} \frac{1}{k(k-1)(k-2)} \cdot \frac{z^{k-3}}{(k-3)!} \leq \frac{1}{6} - \frac{z^2}{60},$$

Since $k(k-1)(k-2) \geq 120$ for $k \geq 6$, it is enough to show that $\sum_{k\geq 6} \frac{z^{k-3}}{(k-3)!} \leq 20 - 2z^2$. But $\sum_{k\geq 6} \frac{z^{k-3}}{(k-3)!} = \sum_{k\geq 3} \frac{z^k}{k!} = \exp(z) - 1 - z - z^2/2$, so we just need to show that $\exp(z) \leq 21 + z - 3z^2/2$ for $z \in (0, 2]$, which is easy to show as the rhs is concave, decreasing for $z \geq 1/3$ and intersecting exp for $z \geq 5/2$. So (22) holds. Since $\log(z) \leq z - 1$, we get

$$\left(\frac{\exp(z) - \exp(-z)}{2z}\right) \log\left(\frac{\exp(z) - \exp(-z)}{2z}\right) \leq \frac{z\exp(z) - z\exp(-z)}{8}$$

Now, we have $\exp(z) - \exp(-z) - 4z \leq 0$ for $z \in [0, 2]$, since the function is strictly convex for $z \geq 0$ with two roots at $z = 0$ and $z > 2$. Reorganising, this shows that $(z\exp(z) - z\exp(-z))/8 \leq z^2/2$ for $z \in [0, 2]$, and so

$$\forall z \in [0, 2] : \left(\frac{\exp(z) - \exp(-z)}{2z}\right) \log\left(\frac{\exp(z) - \exp(-z)}{2z}\right) \leq \frac{z^2}{2}.$$

Putting it together with (21), we now have

$$\begin{aligned}
&D_\varphi\left(z \middle\| (\varphi^*)'\left(\frac{\varphi(z) - \varphi(-z)}{2z}\right)\right) \\
&\quad = \exp(z)\left(\frac{1}{2} - \frac{1}{2z}\right) + \exp(-z)\left(\frac{1}{2} + \frac{1}{2z}\right) \\
&\qquad\qquad + \left(\frac{\exp(z) - \exp(-z)}{2z}\right) \log\left(\frac{\exp(z) - \exp(-z)}{2z}\right) \\
&\quad = \frac{z^2}{2} + \frac{z^2}{2} \\
&\quad = z^2
\end{aligned}$$

for $z \in [0, 2]$, and therefore, since both functions are even, the same holds for $z \in [-2, 0]$ and completes the proof. ■

To show (20), we therefore just need to ensure $c^\star$ small enough so that

$$c^{\star 2} \leq \exp\left(\frac{\mu_P c^\star}{4}\right) - \exp(-\nu_{Q_{t-1}} c^\star). \tag{23}$$

Because $\exp(-\nu_{Q_{t-1}} c^\star)$ is convex, it is upper-bounded over the interval $[0, 2]$ by its chord between its two points in abscissae 0 and 2,

$$\forall c^\star \in [0, 2] : \exp(-\nu_{Q_{t-1}} c^\star) \leq 1 - \frac{1 - \exp(-2\nu_{Q_{t-1}})}{2} c^\star,$$

and we also have, since $\exp(z) \geq 1 + z$,

$$\exp\left(\frac{\mu_P c^\star}{4}\right) \geq 1 + \frac{\mu_P c^\star}{4}.$$

To ensure (23), it is therefore sufficient, as long as $c^\star \in (0, 2]$, that

$$c^{\star 2} \leq \left(\frac{\mu_P}{4} + \frac{1 - \exp(-2\nu_{Q_{t-1}})}{2} \right) c^\star,$$

which, after simplification and considering WLA, is achieved provided

$$c^\star \leq \left(\frac{\gamma_P}{4} + \frac{1 - \exp(-2\gamma_Q)}{2} \right). \tag{24}$$

There remains to check that the condition of Proposition 11 applies, that is, $c^\star \leq 2$. The maximal value of the rhs in (24), taking into account that $\gamma_P, \gamma_Q \leq 1$, is $1/4 + (1 - \exp(-2))/2 \approx 0.57 < 2$, which shows that the condition of Proposition 11 indeed applies and proves Lemma 10. ■

Theorem 12. *Suppose c_t satisfies WLA. Then there exists a constant $\eta > 0$ such that $\eta \cdot c_t$ satisfies WLA and is PS.*

Proof. Even when better bounds are possible, the combination of Lemma 8 and Lemma 10 show that any c_t satisfying the WLA, positively scaled so that $c^\star \leq \log(2)/2$, still satisfies WLA *and* is PS, as claimed. ■

We shall now prove Theorem 15. The proof mainly consists of two lemmata, one showing that $\mathrm{E}_{Q_{t-1}} \exp(\alpha_t c_t)$ is small, the second one showing, under conditions on c_t, that $\mathrm{E}_{Q_{t-1}} \exp(c_t)$ is conveniently upper-bounded by $\mathrm{E}_{Q_{t-1}} \exp(\alpha_t c_t)$, leading to the theorem.

Lemma 13. *Let $\alpha_t \stackrel{\text{def}}{=} \frac{1}{2c^\star} \log\left(\frac{1+\nu_{Q_{t-1}}}{1-\nu_{Q_{t-1}}}\right)$. Then*

$$\mathrm{E}_{Q_{t-1}} \exp(\alpha_t c_t) \leq \sqrt{1 - \nu^2_{Q_{t-1}}}.$$

Proof. We know (Nock & Nielsen, 2007) that

$$\forall a, b \in [-1, 1] : \; 1 - ab \geq \sqrt{1 - a^2} \exp\left(-\frac{b}{2} \log\left(\frac{1+a}{1-a} \right) \right). \tag{25}$$

Let $a \stackrel{\text{def}}{=} \nu_{Q_{t-1}}$ and $b \stackrel{\text{def}}{=} -c_t/c^\star$, for short. Then we obtain

$$\begin{aligned}
\exp(\alpha_t c_t) &= \exp\left(-\left(-c_t \cdot \frac{1}{2c^\star} \log\left(\frac{1+\nu_{Q_{t-1}}}{1-\nu_{Q_{t-1}}} \right) \right) \right) \\
&\stackrel{(25)}{\leq} \frac{1 + \nu_{Q_{t-1}} \frac{c_t}{c^\star}}{\sqrt{1 - \nu^2_{Q_{t-1}}}} \\
&= \frac{1 - \nu_{Q_{t-1}} \cdot \left(-\frac{c_t}{c^\star}\right)}{\sqrt{1 - \nu^2_{Q_{t-1}}}},
\end{aligned} \tag{26}$$

which implies the lemma. ■

Lemma 14. *Fix any $J \geq 0$. Suppose that the two conditions hold:*

$$\mathrm{E}_{Q_{t-1}} \exp(c_t) \leq \exp\left(\frac{J}{2}\right), \tag{27}$$

$$\nu_{Q_{t-1}} \leq \frac{J}{1+J}. \tag{28}$$

Then,

$$\mathrm{E}_{Q_{t-1}} \exp(c_t) \leq \frac{1}{\sqrt{1-\nu^2_{Q_{t-1}}}} \cdot \mathrm{E}_{Q_{t-1}} \exp(\alpha_t c_t + J).$$

Proof. Jensen's inequality yields

$$\mathrm{E}_{Q_{t-1}} \exp(\alpha_t c_t) \geq \exp\bigl(\mathrm{E}_{Q_{t-1}} \alpha_t c_t\bigr) = \exp\bigl(-\alpha_t c^\star \nu_{Q_{t-1}}\bigr),$$

hence we rather show the stronger statement

$$\mathrm{E}_{Q_{t-1}} \exp(c_t) \leq \frac{1}{\sqrt{1-\nu^2_{Q_{t-1}}}} \cdot \exp\bigl(-\alpha_t c^\star \nu_{Q_{t-1}} + J\bigr).$$

We use two inequalities:

$$\forall z \in [0,1] : \ \frac{2z^2}{1-z} \geq 4 \log \frac{1}{\sqrt{1-z^2}} \geq z \log\left(\frac{1+z}{1-z}\right). \tag{29}$$

Let us summarize these as $A \geq B \geq C$. To first check these inequalities, we remark:

- to check $A \geq B$, we simplify it: it yields equivalently $g_1(z) \stackrel{\text{def}}{=} z^2(1+z) \geq -(1-z^2)\log(1-z^2) \stackrel{\text{def}}{=} g_2(z)$. We then check that $g_2'(z) = 2z(1+\log(1-z^2))$ while $g_1'(z) = 2z(1+3z/2)$. Both derivatives are continuous with the same limit in 0 and it is easy to check that for $z \geq 0$, $g_2'(z) \leq g_1'(z)$. Since $g_1(0) = g_2(0)$, we get $A \geq B$;

- to check $B \geq C$, we simplify it, which yields equivalently $g_3(z) \stackrel{\text{def}}{=} (z-2)\log(1-z) - (z+2)\log(1+z) \geq 0$. We have $g_3''(z) = 4z^2/(z^2-1)^2 \geq 0$, which shows the strict convexity of the function. We also have $g_3'(0) = g_3(0) = 0$, which gives $g_3(z) \geq 0$ for all z and shows $B \geq C$.

With the latter ineq. (29) and the expression of α_t for the regular boosting regime, we get

$$\begin{aligned}
&\frac{1}{\sqrt{1-\nu_{Q_{t-1}}^2}} \cdot \exp\bigl(-\alpha_t c^\star \nu_{Q_{t-1}} + J\bigr) \\
&\qquad = \exp\left(J + \log\left(\frac{1}{\sqrt{1-\nu_{Q_{t-1}}^2}}\right) - \frac{\nu_{Q_{t-1}}}{2}\log\left(\frac{1+\nu_{Q_{t-1}}}{1-\nu_{Q_{t-1}}}\right)\right) \qquad (30)\\
&\qquad \geq \exp\left(J - \frac{\nu_{Q_{t-1}}}{4}\log\left(\frac{1+\nu_{Q_{t-1}}}{1-\nu_{Q_{t-1}}}\right)\right)\\
&\qquad \geq \exp\left(J - \frac{1}{4}\cdot\frac{2\nu_{Q_{t-1}}^2}{1-\nu_{Q_{t-1}}}\right). \qquad (31)
\end{aligned}$$

The last inequality follows from the former ineq. (29). Suppose now that we can ensure

$$\frac{2\nu_{Q_{t-1}}^2}{1-\nu_{Q_{t-1}}} \leq 2J. \qquad (32)$$

It would follow from (31) that

$$\frac{1}{\sqrt{1-\nu_{Q_{t-1}}^2}} \cdot \exp\bigl(-\alpha_t c^\star \nu_{Q_{t-1}} + J\bigr) \geq \exp\left(J - \frac{1}{4}\cdot 2J\right) = \exp\left(\frac{J}{2}\right),$$

and so to prove the lemma, we would just need

$$\mathrm{E}_{Q_{t-1}} \exp(c_t) \leq \exp\left(\frac{J}{2}\right),$$

which is precisely (27). To get (32), we equivalently need $\nu_{Q_{t-1}}^2 + J\nu_{Q_{t-1}} - J \leq 0$, that is,

$$\nu_{Q_{t-1}} \leq \frac{1}{2}\cdot(-J + \sqrt{J^2+4J}) \qquad (33)$$

To prove a simpler equivalent condition, we let $g_4(z) \stackrel{\text{def}}{=} (1+z)\sqrt{z^2+4z}/(z(3+z))$. We easily get $\lim_{\downarrow 0} g_4(z) = +\infty$, $\lim_{+\infty} g_4(z) = 1$ and $g_4'(z) = -6(z+4)/N$ with $N \stackrel{\text{def}}{=} (z^2+4z)^{3/2}(3+z)^2 \geq 0$, so $g_4(z) \geq 1$ for all $z \geq 0$, and reordering this inequality yields equivalently $z/(1+z) \leq (1/2)\cdot(-z+\sqrt{z^2+4z})$ for $z \geq 0$, so to get (33), we just require $\nu_{Q_{t-1}} \leq J/(1+J)$, which is (28), and ends the proof of Lemma 14. ■

Let $\alpha_t \stackrel{\text{def}}{=} \min\left\{1, \frac{1}{2c^\star}\log\left(\frac{1+\nu_{Q_{t-1}}}{1-\nu_{Q_{t-1}}}\right)\right\}$. Because there are two regimes for α_t, we define two boosting regimes, a *high boosting regime*, $\alpha_t = 1$ ("clamped"), and a *regular boosting regime*, $\alpha_t < 1$ ("not clamped"). We show two rates of decrease for the KL divergence, one for each regime.

Convergence in the regular boosting regime The WLA alone is sufficient to guarantee a significant decrease of the KL divergence of P from Q_{t-1} at each boosting iteration. The proof of the theorem uses a simple reverse of Jensen's inequality which may be of independent interest. Note that even when we require that c_t meet WLA, the decrease of the KL divergence uses its *actual* values for $\mu_P, \nu_{Q_{t-1}}$, which can yield a substantially larger KL decrease.

Theorem 15. *In the regular boosting regime and under WLA,*

$$\mathrm{KL}(P, Q_t|_{\alpha_t}) \leq \mathrm{KL}(P, Q_{t-1}) - \frac{\mu_P}{4} \log\left(\frac{1+\nu_{Q_{t-1}}}{1-\nu_{Q_{t-1}}}\right).$$

Proof. We have

$$\mathrm{E}_P\, \varepsilon_t = \mathrm{E}_{Q_{t-1}}\left[\frac{\mathrm{d}P}{\mathrm{d}Q_{t-1}} \cdot \varepsilon_t\right] = \mathrm{E}_{Q_{t-1}} \exp(c_t). \tag{34}$$

Hence, combining successively the statements of Lemma 14 (we check below that the conditions of the lemma are indeed satisfied) and Lemma 13, we get:

$$\begin{aligned}
\log \mathrm{E}_P\, \varepsilon_t &= \log \mathrm{E}_{Q_{t-1}} \exp(c_t) \\
&\leq \log\left(\frac{1}{\sqrt{1-\nu^2_{Q_{t-1}}}} \cdot \mathrm{E}_{Q_{t-1}} \exp(\alpha_t c_t + J)\right) \qquad (35) \\
&= \log\left(\frac{\mathrm{E}_{Q_{t-1}} \exp(\alpha_t c_t)}{\sqrt{1-\nu^2_{Q_{t-1}}}} \cdot \exp(J)\right) \\
&\leq \log \exp(J) \\
&= J. \qquad (36)
\end{aligned}$$

On the other hand, WLA yields

$$\mu_P c^\star = \mathrm{E}_P\, c_t = \mathrm{E}_P \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q_{t-1}} \cdot \varepsilon_t\right) = \mathrm{KL}(P, Q_{t-1}) + \mathrm{E}_P \log \varepsilon_t. \tag{37}$$

Since $\alpha_t \geq 0$, it follows from Lemma 5 and (37), (36) in this order that

$$\begin{aligned}
\mathrm{KL}(P, Q_t) &\leq \mathrm{KL}(P, Q_{t-1}) - \alpha_t(\mathrm{KL}(P, Q_{t-1}) + \mathrm{E}_P \log \varepsilon_t) + \alpha_t \log \mathrm{E}_P\, \varepsilon_t \qquad (38) \\
&= \mathrm{KL}(P, Q_{t-1}) - \alpha_t \mu_P c^\star + \alpha_t \log \mathrm{E}_P\, \varepsilon_t \\
&\leq \mathrm{KL}(P, Q_{t-1}) - \alpha_t \mu_P c^\star + \alpha_t J \\
&= \mathrm{KL}(P, Q_{t-1}) - \alpha_t(\mu_P c^\star - J). \qquad (39)
\end{aligned}$$

It remains to fix $J \stackrel{\text{def}}{=} \mu_P c^\star/2$, and we get

$$\begin{aligned}
\mathrm{KL}(P, Q_t) &\leq \mathrm{KL}(P, Q_{t-1}) - \frac{\alpha_t \mu_P c^\star}{2} \qquad (40) \\
&= \mathrm{KL}(P, Q_{t-1}) - \frac{\mu_P}{4} \log\left(\frac{1+\nu_{Q_{t-1}}}{1-\nu_{Q_{t-1}}}\right),
\end{aligned}$$

which is the statement of the theorem. We end up the proof of Theorem 15 by showing that the PS property for c_t implies that the conditions of Lemma 14 are satisfied — hence, Theorem 15 is shown for c_t being PS, which we recall is always possible from Theorem 12 when c_t satisfies the WLA. While it is clear that (27) is one of the PS properties for c_t, we still need to show that the PS ensures (28) with $J = \mu_P c^\star/2$, that is, we need to show that

$$\nu_{Q_{t-1}} \leq \frac{\mu_P c^\star}{2 + \mu_P c^\star}. \tag{41}$$

Recall that we are in the regular boosting regime where we do not clamp α_t, and therefore, if we let

$$\nu_{c^\star} \stackrel{\text{def}}{=} \frac{\exp(2c^\star) - 1}{\exp(2c^\star) + 1} \in (0, 1), \tag{42}$$

then we know that $\nu_{Q_{t-1}} \leq \nu_{c^\star}$, so to have (41), it suffices to ensure $\nu_{c^\star} \leq \mu_P c^\star/(2+\mu_P c^\star)$, which equivalently yields

$$\exp(2c^\star) \leq 2 + \mu_P c^\star,$$

which is the first PS property. This ends the proof of Theorem 15. ■

D.0.2. Proof of Theorem 17

Convergence in the high boosting regime This is where things get interesting; when α_t is clamped to 1, the decrease in the KL divergence at each iteration is *guaranteed* to be of order $c^\star$, and can even be significantly larger depending on the actual values of $\nu_{Q_{t-1}}$ and $\nu_{c^\star}$, defined as

$$\nu_{c^\star} \stackrel{\text{def}}{=} \frac{\exp(2c^\star) - 1}{\exp(2c^\star) + 1} \in (0, 1). \tag{43}$$

Because $\alpha_t = 1$, we have $\nu_{Q_{t-1}} \geq \nu_{c^\star}$, so let us write $\nu_{Q_{t-1}} = (1 + \delta_{t-1})\nu_{c^\star}$ for some $\delta_{t-1} \geq 0$. Note that Theorem 17 doese not assume WLA. It is worthwhile remarking that Theorem 18 is a direct consequence of Theorem 15 above.

We follow some of the same steps as for Theorem 15.

Lemma 16. *Let* $\alpha_t \stackrel{\text{def}}{=} 1$. *Then*

$$\mathrm{E}_{Q_{t-1}} \exp(c_t) \leq \frac{1 - \nu_{Q_{t-1}}\nu_{c^\star}}{\sqrt{1 - \nu_{c^\star}^2}},$$

where $\nu_{c^\star}$ *is defined in* (42).

Proof. We have this time $\mathrm{E}_{Q_{t-1}} \exp(c_t) = \mathrm{E}_{Q_{t-1}} \exp(\alpha_t c_t)$. We use again (25) with $a = \nu_{c^\star}$ and get, instead of (26):

$$\exp(\alpha_t c_t) \leq \frac{1 - \nu_{c^\star} \cdot \left(-\frac{c_t}{c^\star}\right)}{\sqrt{1 - \nu_{c^\star}^2}},$$

which implies the lemma after taking the expectation and remarking that for the choice $a = \nu_{c^\star}, \alpha_t = 1$. ■

Theorem 17. *In the high boosting regime,*

$$\mathrm{KL}(P, Q_t|_{\alpha_t}) \leq \mathrm{KL}(P, Q_{t-1}) - \mu_P c^\star - \nu_{c^\star}^2 \cdot \left(\frac{1}{2} + \frac{\delta_{t-1}}{1-\nu_{c^\star}^2}\right).$$

Proof. Since we get a direct bound on $\mathrm{E}_{Q_{t-1}} \exp(c_t)$, we can achieve the proof of Theorem 17 via (34) and (38) as

$$\begin{aligned}
\mathrm{KL}(P, Q_t) &\leq \mathrm{KL}(P, Q_{t-1}) - \alpha_t(\mathrm{KL}(P, Q_{t-1}) + \mathrm{E}_P \log \varepsilon_t) + \alpha_t \log \mathrm{E}_P\, \varepsilon_t \\
&\leq \mathrm{KL}(P, Q_{t-1}) - \mu_P c^\star + \log \mathrm{E}_P\, \varepsilon_t \\
&\leq \mathrm{KL}(P, Q_{t-1}) - \mu_P c^\star + \log \frac{1 - \nu_{Q_{t-1}} \nu_{c^\star}}{\sqrt{1-\nu_{c^\star}^2}} \\
&= \mathrm{KL}(P, Q_{t-1}) - \mu_P c^\star \\
&\qquad + \log\left(\frac{1-\nu_{c^\star}^2}{\sqrt{1-\nu_{c^\star}^2}} - (\nu_{Q_{t-1}} - \nu_{c^\star}) \cdot \frac{\nu_{c^\star}}{\sqrt{1-\nu_{c^\star}^2}}\right) \\
&= \mathrm{KL}(P, Q_{t-1}) - \mu_P c^\star \\
&\qquad + \log\left(\sqrt{1-\nu_{c^\star}^2} - (\nu_{Q_{t-1}} - \nu_{c^\star}) \cdot \frac{\nu_{c^\star}}{\sqrt{1-\nu_{c^\star}^2}}\right) \\
&= \mathrm{KL}(P, Q_{t-1}) - \mu_P c^\star + \frac{1}{2} \cdot \log(1-\nu_{c^\star}^2) \\
&\qquad + \log\left(1 - (\nu_{Q_{t-1}} - \nu_{c^\star}) \cdot \frac{\nu_{c^\star}}{1-\nu_{c^\star}^2}\right) \\
&\leq \mathrm{KL}(P, Q_{t-1}) - \mu_P c^\star - \frac{\nu_{c^\star}^2}{2} - (\nu_{Q_{t-1}} - \nu_{c^\star}) \cdot \frac{\nu_{c^\star}}{1-\nu_{c^\star}^2} \qquad (44) \\
&\leq \mathrm{KL}(P, Q_{t-1}) - \mu_P c^\star - \nu_{c^\star}^2 \cdot \left(\frac{1}{2} + \frac{\delta_{t-1}}{1-\nu_{c^\star}^2}\right),
\end{aligned}$$

where we have let $\nu_{Q_{t-1}} = (1+\delta_{t-1})\nu_{c^\star}$. In (44), we have used $\log(1-x) \leq -x$. ■

I

Theorem 18. *Suppose WLA holds at each iteration. Then using Q_t as in (4) and α_t as in (9), we are guaranteed that $\mathrm{KL}(P, Q_T) \leq \varrho$ after a number of iterations T satisfying:*

$$T \geq 2 \cdot \frac{\mathrm{KL}(P, Q_0) - \varrho}{\gamma_P \gamma_Q}.$$

Proof. The proof stems from the regular boosting regime, using $\log((1+z)/(1-z)) \geq 2z$ for $z \geq 0$. Better rates are possible using the high boosting regime, and in any case, Q_t as in (4) and α_t as in (9) define a simple boosting algorithm to come up with an analytical expression for Q_T that provably converges to P. ■

D.1. Proof of Theorem 19

We reformulate the theorem involving a new notation for readability purpose in the proof.

Theorem 19. *Suppose WLA and WDA hold at each boosting iteration. Then after T boosting iterations:*

$$\mathrm{KL}(P, Q_T) \leq \left(1 - \frac{\gamma_P \min\{2, \gamma_Q/c^\star\}}{2(1+\Gamma_\varepsilon)}\right)^T \cdot \mathrm{KL}(P, Q_0).$$

Proof. We proceed in two steps, first showing how WDA bounds $\mathrm{KL}(P, Q_{t-1})$. We have by definition $\log(\mathrm{d}P/\,\mathrm{d}Q_{t-1}) + \log \varepsilon_t = c_t \leq c^\star$, and so, taking expectations, we get $\mathrm{KL}(P, Q_{t-1}) + c^\star \mu_{\varepsilon_t} \leq \int \mathrm{d}P c^\star = c^\star$. Hence,

$$\mathrm{KL}(P, Q_{t-1}) \leq c^\star - c^\star \mu_{\varepsilon_t} \leq (1+\Gamma_\varepsilon) c^\star.$$

We now show the statement of the theorem. Suppose we are in the low-boosting regime where α_t is not clamped. In this case, since $\log((1+z)/(1-z)) \geq 2z$, we have

$$\alpha_t \geq \frac{\nu_{Q_{t-1}}}{c^\star} \geq \gamma_r,$$

and it comes from (40)

$$\mathrm{KL}(P, Q_t) \leq \mathrm{KL}(P, Q_{t-1}) - \frac{\gamma_r \gamma_P c^\star}{2}.$$

In the high-boosting regime, we have immediately $\mathrm{KL}(P, Q_t) \leq \mathrm{KL}(P, Q_{t-1}) - \gamma_P c^\star$. So, letting $\rho \stackrel{\text{def}}{=} \min\{1, \gamma_r/2\}$, we get under the assumptions of the theorem $\mathrm{KL}(P, Q_t) \leq \mathrm{KL}(P, Q_{t-1}) - \rho \gamma_P c^\star$, and WDA yields in addition through D.1,

$$\begin{aligned}
\mathrm{KL}(P, Q_t) &\leq \mathrm{KL}(P, Q_{t-1}) - \frac{\rho \gamma_P}{1+\Gamma_\varepsilon} \cdot \mathrm{KL}(P, Q_{t-1}) \\
&= \left(1 - \frac{\min\{1, \gamma_r/2\}\gamma_P}{1+\Gamma_\varepsilon}\right) \cdot \mathrm{KL}(P, Q_{t-1}) \\
&= \left(1 - \frac{\min\{2, \gamma_r\}\gamma_P}{2(1+\Gamma_\varepsilon)}\right) \cdot \mathrm{KL}(P, Q_{t-1}),
\end{aligned}$$

and we get the statement of the theorem by replacing γ_r by its expression, completing the proof ■

D.1.1. Proof of Theorem 23

The proof of Theorem 23 is essentially a rewriting of the proof of Theorem 15 and Theorem 17, taking into account that we have just samples from distributions to compute the estimates of edges and WLA. We split the proof in three steps, one that provides an additional Lemma we shall need for the next steps, one for the non-clamped regime for α_t, one for the clamped regime for α_t.

Step.1, We need the additional simple lemma, which is an exploitation of basic concentration inequalities (McDiarmid, 1998, §3.1).

Lemma 20. *For any $0 < \delta \leq 1$ and $0 < \kappa \leq 1$, suppose the weak learner samples at each iteration $t = 1, 2, ..., T$, m_P times P and m_Q times Q_t, such that the following constraints hold:*

$$m_P \geq \frac{1}{\kappa^2 {\gamma_P}^2} \log \frac{4T}{\delta} \quad \textit{and} \quad m_Q \geq \frac{1}{\kappa^2 {\gamma_Q}^2} \log \frac{4T}{\delta}.$$

Then there is probability $\geq 1 - \delta$ that for any $t = 1, 2, ..., T$, the current estimators $\hat{\mu}_P$ of μ_P and $\hat{\nu}_{Q_{t-1}}$ of $\nu_{Q_{t-1}}$ satisfy:

$$|\hat{\mu}_P - \mu_P| \leq \kappa \gamma_P, \tag{45}$$
$$|\hat{\nu}_{Q_{t-1}} - \nu_{Q_{t-1}}| \leq \kappa \gamma_Q. \tag{46}$$

From now on, we denote as E the proposition that both (45) and (46) hold for all T iterations, for some $0 < \kappa \leq 1$ that will be computed later.

We have a slightly weaker version of Lemma 13, straightforward to prove from Lemma 13.

Lemma 21. *Let $\alpha_t \stackrel{\text{def}}{=} \frac{1}{2c^\star} \log\left(\frac{1+\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right)$. Then we have under E,*

$$\mathrm{E}_{Q_{t-1}} \exp(\alpha_t c_t) \leq \sqrt{1 - \hat{\nu}^2_{Q_{t-1}}} + \frac{\kappa \gamma_Q \hat{\nu}_{Q_{t-1}}}{\sqrt{1 - \hat{\nu}^2_{Q_{t-1}}}}.$$

Lemma 22. *Fix any $J \geq 0$. Suppose that the two conditions hold:*

$$\mathrm{E}_{Q_{t-1}} \exp(c_t) \leq \exp\left(\frac{J}{2}\right), \tag{47}$$
$$\hat{\nu}_{Q_{t-1}} \leq \frac{J}{1+J}. \tag{48}$$

Then we have under E,

$$\mathrm{E}_{Q_{t-1}} \exp(c_t) \leq \frac{1}{\sqrt{1 - \hat{\nu}^2_{Q_{t-1}}}} \cdot \mathrm{E}_{Q_{t-1}} \exp(\alpha_t c_t + J) \cdot \exp\left(\frac{\kappa \gamma_Q}{2} \log\left(\frac{1 + \hat{\nu}_{Q_{t-1}}}{1 - \hat{\nu}_{Q_{t-1}}}\right)\right).$$

Proof. Because the proof mixes the use of $\hat{\nu}_{Q_{t-1}}$ and $\nu_{Q_{t-1}}$, we re-sketch the major lines of the proof from Lemma 14. First, Jensen's inequality still yields $\mathrm{E}_{Q_{t-1}} \exp(\alpha_t c_t) \geq \exp\big(-\alpha_t c^\star \nu_{Q_{t-1}}\big)$, so we in fact prove

$$\frac{1}{\sqrt{1 - \hat{\nu}^2_{Q_{t-1}}}} \cdot \exp\big(-\alpha_t c^\star \nu_{Q_{t-1}} + J\big) \geq \mathrm{E}_{Q_{t-1}} \exp(c_t).$$

The chain of (in)equalities in (30)–(31) now becomes withe the use of E:

$$\begin{aligned}
&\frac{1}{\sqrt{1-\hat{\nu}_{Q_{t-1}}^2}} \cdot \exp\big(-\alpha_t c^\star \nu_{Q_{t-1}} + J\big) \\
&\qquad = \exp\left(J + \log\left(\frac{1}{\sqrt{1-\hat{\nu}_{Q_{t-1}}^2}}\right) - \frac{\nu_{Q_{t-1}}}{2}\log\left(\frac{1+\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right)\right) \\
&\qquad \geq \exp\left(J + \log\left(\frac{1}{\sqrt{1-\hat{\nu}_{Q_{t-1}}^2}}\right) - \frac{\hat{\nu}_{Q_{t-1}}}{2}\log\left(\frac{1+\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right)\right. \\
&\qquad\qquad \left. - \frac{\kappa\gamma_Q}{2}\log\left(\frac{1+\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right)\right) \\
&\qquad \geq \exp\left(J - \frac{1}{4}\cdot\frac{2\hat{\nu}_{Q_{t-1}}^2}{1-\hat{\nu}_{Q_{t-1}}} - \frac{\kappa\gamma_Q}{2}\log\left(\frac{1+\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right)\right).
\end{aligned}$$

Provided we have $\hat{\nu}_{Q_{t-1}} \leq J/(1+J)$, which is (48), we have similarly to Lemma 14,

$$\frac{2\hat{\nu}_{Q_{t-1}}^2}{1-\hat{\nu}_{Q_{t-1}}} \leq 2J.$$

Hence, it follows that

$$\begin{aligned}
&\exp\left(\frac{\kappa\gamma_Q}{2}\log\left(\frac{1+\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right)\right) \cdot \frac{1}{\sqrt{1-\nu_{Q_{t-1}}^2}} \cdot \exp\big(-\alpha_t c^\star \nu_{Q_{t-1}} + J\big) \\
&\quad \geq \exp\left(J - \frac{1}{4}\cdot 2J\right) \\
&\quad = \exp\left(\frac{J}{2}\right),
\end{aligned}$$

and so to prove the lemma, we would just need

$$\mathrm{E}_{Q_{t-1}} \exp(c_t) \leq \exp\left(\frac{J}{2}\right),$$

which is (47). ■

Now, instead of (35)–(36), we get

$$\begin{aligned}
\log \mathrm{E}_P\, \varepsilon_t &\leq \log\left(\frac{1}{\sqrt{1-\nu_{Q_{t-1}}^2}} \cdot \mathrm{E}_{Q_{t-1}} \exp(\alpha_t c_t + J)\right) + \frac{\kappa\gamma_Q}{2} \log\left(\frac{1+\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right) \\
&= \log\left(\frac{\mathrm{E}_{Q_{t-1}} \exp(\alpha_t c_t)}{\sqrt{1-\nu_{Q_{t-1}}^2}} \cdot \exp(J)\right) + \frac{\kappa\gamma_Q}{2} \log\left(\frac{1+\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right) \\
&\leq \log\left(\left(1 + \frac{\kappa\gamma_Q \hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}^2}\right)\exp(J)\right) + \frac{\gamma_Q}{4} \log\left(\frac{1+\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right) \\
&= J + \log\left(1 + \frac{\kappa\gamma_Q \hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}^2}\right) + \frac{\kappa\gamma_Q}{2} \log\left(\frac{1+\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right).
\end{aligned}$$

We get from (39)

$$\mathrm{KL}(P, Q_t) \leq \mathrm{KL}(P, Q_{t-1}) - \alpha_t(\mu_P c^\star - J - J') \tag{49}$$

with, because $\log(1+x) \leq x$,

$$\begin{aligned}
J' &\stackrel{\text{def}}{=} \log\left(1 + \frac{\kappa\gamma_Q \hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}^2}\right) + \frac{\kappa\gamma_Q}{2} \log\left(\frac{1+\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right) \\
&= \log\left(1 + \frac{\kappa\gamma_Q \hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}^2}\right) + \frac{\kappa\gamma_Q}{2} \log\left(1 + \frac{2\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right) \\
&\leq \frac{\kappa\gamma_Q \hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}^2} + \frac{\kappa\gamma_Q \hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}} \\
&\leq \kappa \cdot \frac{2\gamma_Q \hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}
\end{aligned}$$

Now, we would like from the PS property and (41) that we have:

$$\hat{\nu}_{Q_{t-1}} \leq \frac{\mu_P c^\star}{2 + \mu_P c^\star}, \tag{50}$$

so

$$J' \leq \kappa\gamma_Q \mu_P c^\star,$$

and we get from (49),

$$\mathrm{KL}(P, Q_t) \leq \mathrm{KL}(P, Q_{t-1}) - \alpha_t((1-\kappa\gamma_Q)\mu_P c^\star - J),$$

and if we fix again $J = \mu_P c^\star/2$, we get this time

$$\mathrm{KL}(P, Q_t) \leq \mathrm{KL}(P, Q_{t-1}) - \alpha_t \cdot \left(\frac{1}{2} - \kappa\gamma_Q\right) \cdot \mu_P c^\star.$$

If we pick κ satisfying

$$\kappa \leq \min\left\{1, \frac{1}{4\gamma_Q}\right\}, \tag{51}$$

then we are guaranteed $1/2 - \kappa\gamma_Q \geq 1/4$ and so

$$\mathrm{KL}(P, Q_t) \leq \mathrm{KL}(P, Q_{t-1}) - \frac{\mu_P}{8} \log\left(\frac{1+\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right), \tag{52}$$

In the same way as for Theorem 15, we ensure (50) by noting that, since we are in the case where we do not clamp α_t, letting

$$\hat{\mu}_{c^\star} \stackrel{\text{def}}{=} \frac{\exp(2c^\star) - 1}{\exp(2c^\star) + 1} \in (0, 1),$$

then we again need to ensure $\nu_{c^\star} \leq \mu_P c^\star/(2 + \mu_P c^\star)$, which again yields to the first PS property.

We are not yet done as we now have to replace μ_P by its estimate, $\hat{\mu}_P$, in (52). Under E, we obtain

$$\mathrm{KL}(P, Q_t) \leq \mathrm{KL}(P, Q_{t-1}) - \frac{\hat{\mu}_P - \kappa\gamma_P}{8} \log\left(\frac{1+\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right),$$

and under the (EWLA), we know that $\hat{\mu}_P \geq \gamma_P$, so if we also put the constraint $\kappa \leq 1/2$, then $\kappa\gamma_P \leq \gamma_P/2 \leq \hat{\mu}_P/2$ and so:

$$\mathrm{KL}(P, Q_t) \leq \mathrm{KL}(P, Q_{t-1}) - \frac{\hat{\mu}_P}{16} \log\left(\frac{1+\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right),$$

as claimed. This ends the proof of Step.2 by remarking two additional facts: (i) we have not changed the PS properties, and (ii) we have two constraints over κ (also considering (51)), which can be both satisfied by choosing (since $\gamma_Q \leq 1$) κ satisfying

$$\kappa \leq \frac{1}{4}. \tag{53}$$

Theorem 23. *Suppose* $\mathrm{EWLA}_{\delta,T}$ *holds. Then with probability of at least* $1 - \delta$*,*

$$\forall t = 1, 2, ..., T: \ \mathrm{KL}(P, Q_t) \leq \mathrm{KL}(P, Q_{t-1}) - \Delta_t,$$

where

$$\Delta_t \stackrel{\text{def}}{=} \begin{cases} \frac{\hat{\mu}_P}{16} \log\left(\frac{1+\hat{\nu}_{Q_{t-1}}}{1-\hat{\nu}_{Q_{t-1}}}\right) & \textit{in the non-clamped regime,} \\ \frac{\hat{\mu}_P c^\star}{2} + \nu_{c^\star}^2 \cdot \left(\frac{1}{4} + \frac{\hat{\delta}_{t-1}}{1-\nu_{c^\star}^2}\right) & \textit{otherwise.} \end{cases}$$

Proof. We proceed in exactly the same way as we did for Theorem 17. We first remark that Lemma 21 *is still valid* in this case, so that we still have

$$\mathrm{E}_{Q_{t-1}} \exp(c_t) \leq \frac{1 - \nu_{Q_{t-1}} \nu_{c^\star}}{\sqrt{1 - \nu_{c^\star}^2}}.$$

It is not hard to check that we then keep the exact same derivations as for Theorem 17, yielding

$$\mathrm{KL}(P, Q_t) \leq \mathrm{KL}(P, Q_{t-1}) - \mu_P c^\star - \nu_{c^\star}^2 \cdot \left(\frac{1}{2} + \frac{\delta_{t-1}}{1 - \nu_{c^\star}^2} \right),$$

where we have let $\nu_{Q_{t-1}} = (1 + \delta_{t-1}) \nu_{c^\star}$. Remark that this time, δ_{t-1} is not necessarily positive since we do not have access to $\nu_{Q_{t-1}}$ — this may happen when $\nu_{Q_{t-1}} < \hat{\nu}_{Q_{t-1}}$. What we do, to finish up Step.3, is replace δ_{t-1} by the $\hat{\delta}_{t-1}$ for which we have $\hat{\nu}_{Q_{t-1}} = (1 + \hat{\delta}_{t-1}) \nu_{c^\star}$, which we are then sure is going to satisfy $\hat{\delta}_{t-1} \geq 0$ under the clamped regime for α_t. Under E, we have

$$\begin{aligned}
\delta_{t-1} &= \frac{\nu_{Q_{t-1}}}{\nu_{c^\star}} - 1 \\
&\geq \frac{\hat{\nu}_{Q_{t-1}}}{\nu_{c^\star}} - 1 - \kappa \cdot \frac{\gamma_Q}{\nu_{c^\star}} \\
&= \hat{\delta}_{t-1} - \kappa \cdot \frac{\gamma_Q}{\nu_{c^\star}}
\end{aligned}$$

yielding

$$\mathrm{KL}(P, Q_t) \leq \mathrm{KL}(P, Q_{t-1}) - \mu_P c^\star - \nu_{c^\star}^2 \cdot \left(\frac{1}{2} - \kappa \cdot \frac{\gamma_Q}{\nu_{c^\star}(1 - \nu_{c^\star}^2)} + \frac{\hat{\delta}_{t-1}}{1 - \nu_{c^\star}^2} \right),$$

Suppose we pick κ such that

$$\kappa \leq \frac{\nu_{c^\star}(1 - \nu_{c^\star})}{2}. \tag{54}$$

Since $\nu_{c^\star} \in [0, 1]$, we also have

$$\kappa \leq \frac{\nu_{c^\star}(1 - \nu_{c^\star}^2)}{2}.$$

In this case, we obtain, since $\gamma_Q \leq 1$,

$$\mathrm{KL}(P, Q_t) \leq \mathrm{KL}(P, Q_{t-1}) - \mu_P c^\star - \nu_{c^\star}^2 \cdot \left(\frac{1}{4} + \frac{\hat{\delta}_{t-1}}{1 - \nu_{c^\star}^2} \right).$$

Finally, we also know under E that $\mu_P c^\star \geq \hat{\mu}_P c^\star - \kappa \gamma_P c^\star$. Under the (EWLA), we know that $\hat{\mu}_P \geq \gamma_P$, so if we again put the constraint $\kappa \leq 1/2$ (satisfied from (53)), then $\kappa \gamma_P c^\star \leq \gamma_P c^\star / 2 \leq \hat{\mu}_P c^\star / 2$ and so:

$$\mathrm{KL}(P, Q_t) \leq \mathrm{KL}(P, Q_{t-1}) - \frac{\hat{\mu}_P c^\star}{2} - \nu_{c^\star}^2 \cdot \left(\frac{1}{4} + \frac{\hat{\delta}_{t-1}}{1 - \nu_{c^\star}^2} \right),$$

which ends the proof of Step.3 once we remark that (53) and (54) are both satisfied if

$$\kappa = \min\left\{\frac{1}{4}, \frac{\nu_{c^\star}(1-\nu_{c^\star})}{2}\right\} = \frac{\nu_{c^\star}(1-\nu_{c^\star})}{2} = \kappa^\star.$$

■

E. Experimental procedure

All models were trained using the ADAM optimiser with the default settings from FLUX.JL (Innes, 2018): $\eta = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1e-08$. In all experiments we divide the data into training (75%) and test (25%) sets, which we use to early stop on certain experiments. The reset of the experimental conditions are presented in Table 14. Each experiment was run 20 times.

Table 14: Experimental procedure

Experiment	P	Q_0	α_t	Sample size (P, Q)	Epochs	Batch size	Early stop[a]	Network topology of c_t
§4.1.1, §4.1.2	8 mode Gaussian ring mixture $\sigma = 1$	Isotropic Gaussian $\sigma = 1$	1/2	(1000,1000)	3000	50	Not used	$\mathcal{X} \xrightarrow[\text{dense}]{\text{ReLU}} \mathbb{R}^5 \xrightarrow[\text{dense}]{\text{ReLU}} \mathbb{R}^5 \xrightarrow[\text{dense}]{\text{ReLU}} \mathbb{R}$
§4.1.3	8 mode Gaussian ring mixture $\sigma = 1$	Isotropic Gaussian $\sigma = 1$	1/2	(1000,1000)	3000	50	Not used	varies
§4.1.4	randomly arranged 8 mode Gaussian mixture	Isotropic Gaussian $\sigma = 1$	1/2	(1000,1000)	2000	250	20%	$\mathcal{X} \xrightarrow[\text{dense}]{\text{ReLU}} \mathbb{R}^{10} \xrightarrow[\text{dense}]{\text{ReLU}} \mathbb{R}^{10} \xrightarrow[\text{dense}]{\text{ReLU}} \mathbb{R}$
§4.1.6	randomly arranged 8 mode Gaussian mixture	Empirically fit Gaussian	Selected to minimise NLL	(1000,1000)	3000	50	Not used	$\mathcal{X} \xrightarrow[\text{dense}]{\text{ReLU}} \mathbb{R}^{10} \xrightarrow[\text{dense}]{\text{ReLU}} \mathbb{R}^{10} \xrightarrow[\text{dense}]{\text{ReLU}} \mathbb{R}$
§4.1.5	ADAGAN generated randomly arranged 8 mode Gaussian mixture	Empirically fit Gaussian	Selected to minimise NLL	(5000,5000)[b]	1000	250	3%	$\mathcal{X} \xrightarrow[\text{dense}]{\text{ReLU}} \mathbb{R}^{10} \xrightarrow[\text{dense}]{\text{ReLU}} \mathbb{R}^{10} \xrightarrow[\text{dense}]{\text{ReLU}} \mathbb{R}$

[a] This parameter terminates training when the the test error falls this amount below the training error. Useful to stabilise training that might otherwise fail due to exploding test error.

[b] ADAGAN trains on a set of (64000,64000) samples, we take a subset of these to use for training Q_t.