Boosted Density Estimation Remastered

Zac Cranko[1,2] and Richard Nock[2,1,3]

[1]The Australian National University [2]CSIRO Data 61 [3]The University of Sydney

Other approaches

Approach	Updates	Rate	Assumption
Dai et al., 2016	kernel density estimate / particles	$\Omega(\mathrm{KL}(P,Q_0))$	smoothness, Lipschitz, measure concentration, etc.
Tolstikhin et al., 2017	density	$\Omega(\log \mathrm{JS}(P,Q_0))$	updates close to optimal
Grover & Ermon, 2018	density	none	none
This work	binary classifiers	$\Omega(\log \mathrm{KL}(P,Q_0))$	weak learning assumption on classifiers, weak dominance

Variational f-divergences

- $f:\mathbb{R}_+\to\mathbb{R}_+$ is convex
- useful to empirically estimate an f-divergence
- origin of f-GAN
- supremum is achieved for $u = f'\circ \mathrm{d}P/\mathrm{d}Q$

$$\begin{aligned}\mathrm{I}_f(P,Q) &= \int (f^*)^*\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}Q\\ &= \int \sup_{t>0}\left(t\cdot\frac{\mathrm{d}P}{\mathrm{d}Q} - f^*(t)\right)\mathrm{d}Q\\ &= \sup_{u\in(\mathrm{dom}\, f^*)^{\mathcal{X}}}\int\left(u\cdot\frac{\mathrm{d}P}{\mathrm{d}Q} - f^*\circ u\right)\mathrm{d}Q\\ &= \sup_{u\in(\mathrm{dom}\, f^*)^{\mathcal{X}}}\Big(\mathrm{E}_P\, u - \mathrm{E}_Q\, f^*\circ u\Big),\end{aligned}$$

Reparameterisation

	I_f	$f(t)$	$f^*(t^*)$	$f'(t)$	$(f^*\circ f')(t)$
Kullback–Liebler	KL	$t\log t$	$\exp(t^*-1)$	$\log t+1$	t
Reverse KL	rKL	$\lvert t-1\rvert$	$-\log(-t^*)-1$	$-1/t$	$\log t-1$
Hellinger	-	$(\sqrt{t}-1)^2$	$3(t^*-1)^{-1}-1$	$1-1/t$	$\sqrt{t}-1$
Pearson	χ^2	$(t-1)^2$	$t^*(4+t^*)/4$	$2(t-1)$	t^2-1
GAN	GAN	$t\log t-(t+1)\log(t+1)$	$-\log(1-\exp(t^*))$	$-\log(t)-\log(t+1)$	$\log(1+t)$

Fit distributions Q_t using

$$\tilde{Q}_t(\mathrm{d}x) = d_t^{\alpha_t}(x)\cdot\tilde{Q}_{i-1}(\mathrm{d}x),\quad Q_t = \frac{1}{Z_t}\tilde{Q}_t,\quad \text{where}\quad Z_t \stackrel{\mathrm{def}}{=} \int \mathrm{d}\tilde{Q}_t,$$
$$d_t \in \arg\min_d J(d) \stackrel{\mathrm{def}}{=} \mathrm{E}_{Q_{t-1}} f^*\circ f'\circ d - \mathrm{E}_P\, f'\circ d.$$

A canonical activation for f-GAN

Let $\varphi(D) \stackrel{\mathrm{def}}{=} \frac{D}{1-D}$. The GAN objective is implicitly solving the reparameterised variational problem:

$$\begin{aligned}&\sup_{D\in(0,1)^{\mathcal{X}}}\left(\mathrm{E}_P\log(D) + \mathrm{E}_Q\log(1-D)\right)\\ &\quad= \sup_{D\in(0,1)^{\mathcal{X}}}\left(\mathrm{E}_P(f'\circ\varphi)\circ D - \mathrm{E}_Q(f^*\circ f'\circ\varphi)\circ D\right)\\ &\quad= \sup_{d\in(0,\infty)^{\mathcal{X}}}\left(\mathrm{E}_P\, f'\circ d - \mathrm{E}_Q(f^*\circ f')\circ d\right).\end{aligned}$$

The function $f'\circ\varphi$ serves a canonical choice for the activation function g_f of Nowozin et al. (2016).

Assumptions

Let $\varepsilon_t \stackrel{\mathrm{def}}{=} \frac{\mathrm{d}Q_{t-1}}{\mathrm{d}P}\cdot d_t$, and $c^\star \stackrel{\mathrm{def}}{=} \max_t \operatorname{ess\,sup}|c_t|$, and choose $\alpha_t \stackrel{\mathrm{def}}{=} \min\left\{1, \frac{1}{2c^\star}\log\left(\frac{1+\nu_{Q_{t-1}}}{1-\nu_{Q_{t-1}}}\right)\right\}$.

Weak learning assumption (WLA)

$$\nu_{Q_{t-1}} \stackrel{\mathrm{def}}{=} \frac{1}{c^\star}\mathrm{E}_{Q_{t-1}}[-c_t] \quad\text{and}\quad \mu_P \stackrel{\mathrm{def}}{=} \frac{1}{c^\star}\mathrm{E}_P[c_t],$$
$$\exists \gamma_P,\gamma_Q\in(0,1]:\ \mu_P\geq\gamma_P,\quad \nu_{Q_{t-1}}\geq\gamma_Q.$$

Weak dominance assumption (WDA)

$$\mu_{\varepsilon_t} \stackrel{\mathrm{def}}{=} \frac{1}{c^\star}\cdot\mathrm{E}_P\log\varepsilon_t,$$
$$\exists \Gamma_\varepsilon>0, \forall t\geq 1:\ \mu_{\varepsilon_t}\geq -\Gamma_\varepsilon$$

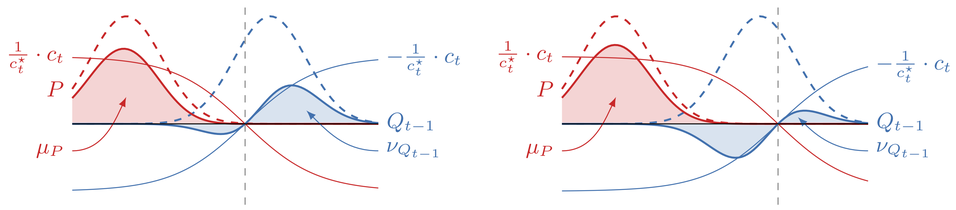
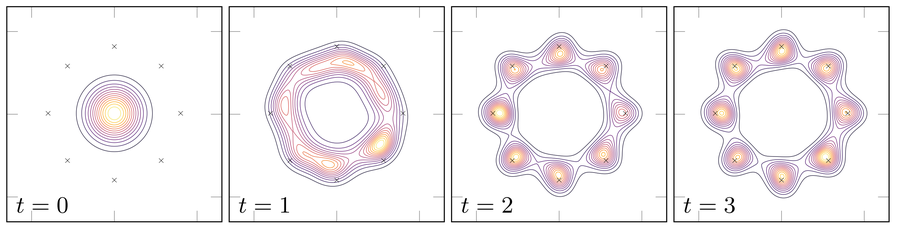


Illustration of WLA being satisfied and violated

Convergence



Convergence

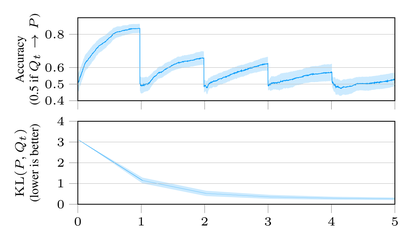
Suppose WLA holds at each iteration. Then we are guaranteed that $\mathrm{KL}(P,Q_T)\leq\varrho$ after a number of iterations T satisfying:

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

Geometric convergence

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

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



As $Q_{t-1}\to P$ it becomes harder to build a classifier to tell them apart.

Reverse Jensen inequality

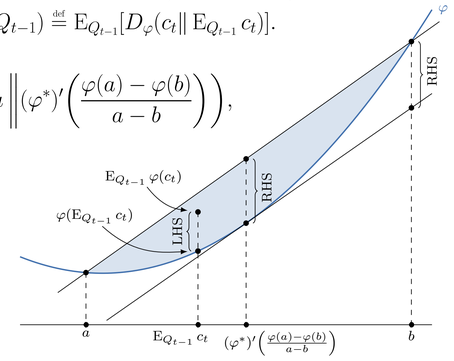
Suppose $\varphi:\mathcal{X}\to\mathbb{R}$ strictly convex differentiable and $c_t(x)\in[a,b]$. The *minimal Bregman information* of (c_t,Q_{t-1}) relative to φ is

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

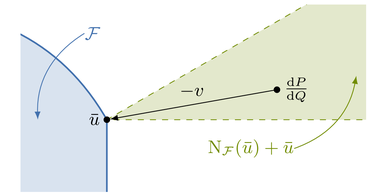
Then

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

where u can be chosen to be a or b.



Inexact variational solutions



Assume $f:\mathbb{R}_+\to\mathbb{R}_+$ is strictly convex and twice differentiable, and $\mathcal{F}$ is a normed space of functions $\mathcal{X}\to\mathrm{int}(\mathrm{dom}\, f)$. Let $\mathcal{F}\subseteq\mathcal{F}$ 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 $\mathrm{d}P/\mathrm{d}Q\in\mathrm{int}\,\mathcal{F}$, then $\bar{u}=\mathrm{d}P/\mathrm{d}Q$.

References

Dai, B., He, N., Dai, H., and Song, L. Provable bayesian inference via particle mirror descent. In *Artificial Intelligence and Statistics*, pp. 985–994, 2016.

Grover, A. and Ermon, S. Boosted generative models. In *AAAI*, 2018.

Nowozin, S., Cseke, B., and Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pp. 271–279, 2016.

Tolstikhin, I.-O., Gelly, S., Bousquet, O., Simon-Gabriel, C., and Schölkopf, B. Adagan: Boosting generative models. In *NIPS*, pp. 5430–5439, 2017.