Power efficient deep learning

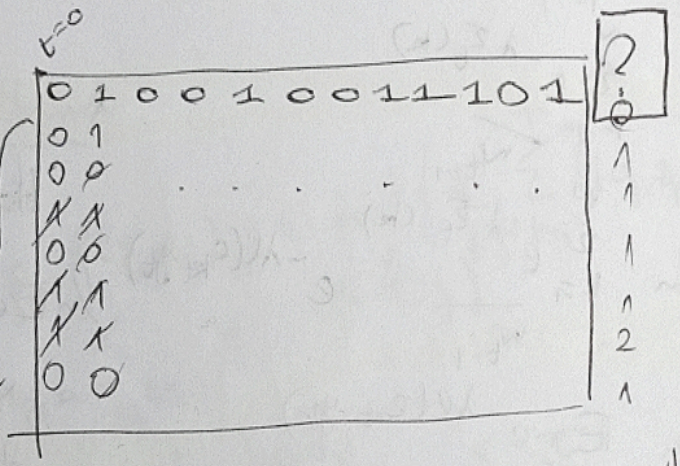
(1) Toy example / finite case.

Obj : introduire de nouveaux outils environnementaux dans l'apprentissage des réseaux de neurones. ⇝ 5 ans.

- non-parametric statistics / tradeoff complexity / fitting the data
- Bayesian priors.
- optimal transport and cost function.

1. online learning & exponential weighted averages.

toy example



⇝ $\hat{y}_t := \bar{y}_t(y_1, \dots y_{t-1})$

$e_{1k}, k = 1, N$,

$e_{2k}, \dots$

hypothesis : 1 expert knows the story.

solution : keep or kill.

$err(T) \leq N - 1$ ⇒ chaque fois on se trompe on tue au hasard mais après $N-1$ erreurs il ne reste que l'oracle.

keep or kill + majority vote optimale. $err(T) \leq \log N$

W_e = nombre de experts après e erreurs.

$W_e \leq \frac{1}{2} W_{e-1}$ (HV).

$W_e \leq 2^{-e} W_0$ $\quad W_0 = N$

$2^e \leq W_0 = N$. $\quad e \leq \log_2 N$.

General case: uncountable $\mathcal{G}$ / Deep learning.

Toy example → general case (finite set of experts) → uncountable set.

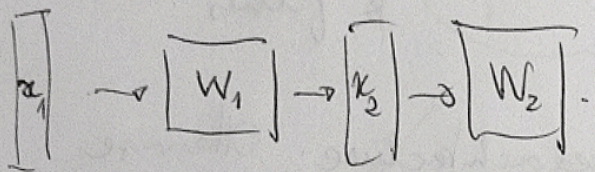
$z_1, \dots, z_t$? $\quad \mathcal{G} = \{ g_\theta, \theta \in \Theta \subseteq \mathbb{R}^p \}$.

↳ deep nets, linear regression, etc

$z_t = (x_t, y_t)$. $\quad g_\theta(x_t) = \langle x_t, \theta \rangle$

$g_\theta(x_t) = \langle x_t, f_\theta \rangle_p$,

$= \sum_{k=1}^{p} \theta_k f_k(x_t)$



$\theta = (W_1, \dots, W_L) \subseteq \mathbb{R}^p$, $\quad \text{span}\{f_1, \dots, f_p\} = \mathcal{G}$.

$p \approx$ 60 millions.

PAC-Bayesian bound.

$$\sum_{t=1}^{T} \mathbb{E}_{g \sim \hat{\rho}_t} \ell(g, z_t) \leq \inf_{\rho \in \mathcal{M}(\mathcal{G})} \left\{ \mathbb{E}_{g \sim \rho} \sum_{t=1}^{T} \ell(g, z_t) + \frac{K(\rho, \pi)}{\lambda} \right\}.$$

$\ell(g, z_t) \leftarrow |y_t - g_{\theta_t}(x_t)|^2$

where $K(\rho, \pi) = \begin{cases} \int \log \frac{d\rho}{d\pi} \, d\rho & \rho \ll \pi \\ +\infty & \text{if } \rho \not\ll \pi \end{cases}$ $\quad (\pi(g) = 0 \Rightarrow \rho(g) = 0.)$

$\hat{\rho}_{t+1} \sim \exp\left(-\sum_{u=1}^{t} \ell(g, z_t) \right) \pi$

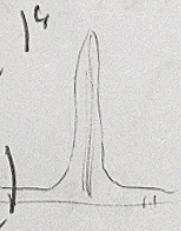
Kullback-Leibler divergence.

power tail.

Proof (sketch). Audibert 2009. Consequence.

sparsity prior. → $\pi(\theta) = \prod_{j=1}^{p} \text{Student}_1(\theta_j) \sim \frac{1}{|\theta_j|} \cdot \frac{\text{const}}{\left(1 + \frac{|x_t|}{\tau}\right)^4}$

$K(\rho_{W_0}, \pi) = \int \rho_{W_0}(w) \log \frac{\rho_{W_0}(w)}{\pi(w)} \, dw = \|W_0\|_0 \log\left(1 + \frac{\|W\|_1}{\|W\|_0 \tau}\right)$



translate de τ de direction w_0. $\quad \rho_{W_0}(w) = \pi(w - w_0)$.

N nombre de coefficients non nuls.

(3) Generalization - generic regret bounds

$$\rho_{t+1} = \arg\min_{\rho} \left\{ \sum_{u=1}^{t} \tilde{\ell}(g_u, x_u) + \frac{K(\rho,\pi)}{\lambda} \right\} = \arg\min_{\rho} \left\{ \ell(g_t, x_t) + \frac{K(\rho,\rho_t)}{\lambda} \right\}$$

convex duality formula

$$\rho(dg) = \exp(-h(g))\,\pi(dg)$$

$$= \arg\min_{\rho} \left\{ \mathbb{E}_{g\sim\rho} h(g) + \frac{K(\rho,\pi)}{\lambda} \right\}$$

generalization

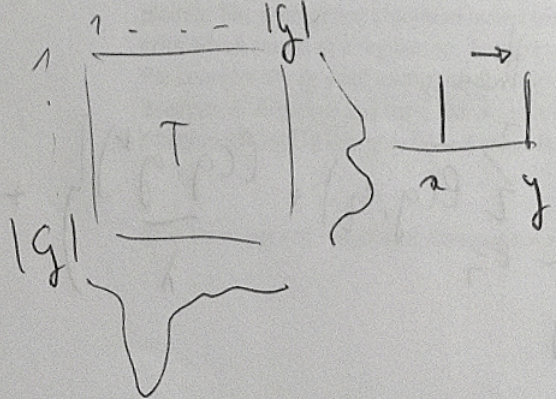
$$\min_{\rho} \left\{ \mathbb{E}_{g\sim\rho} h(g) + \frac{\mathcal{D}(\rho,\pi)}{\lambda} \right\}, \text{ where } \mathcal{D}(\rho,\pi) = \text{Bregman divergence}$$

optimal transport

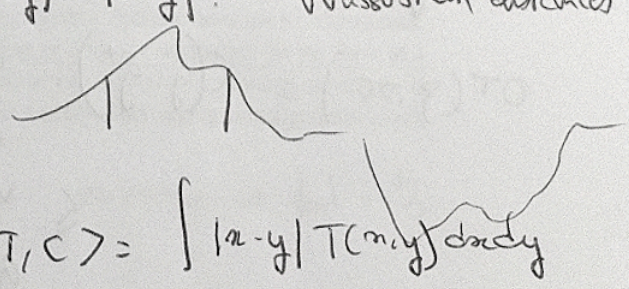
$$B_\phi(\rho,\pi) = \phi(\rho) - \phi(\pi) - \nabla\phi(\pi)(\rho-\pi).$$

$$\mathcal{T}(\rho,\pi) = \min_{T\in\Delta(\rho,\pi)} \langle T, C\rangle \quad \text{where } C: \mathcal{G}\times\mathcal{G}\to\mathbb{R} \text{ cost function to move a unit mass from } g \text{ to } g'.$$

$$\Delta(\rho,\pi) = \left\{ T: \mathcal{G}\times\mathcal{G}\to\mathbb{R}_+ : \sum_{g'} T(g,g') = \rho(g), \sum_{g} T(g,g') = \pi(g') \right\}.$$



$c(x,y) = |x-y|$ Wasserstein distances



$$\langle T, C\rangle = \int |x-y|\, T(x,y)\, dx\, dy$$

$$H(\rho,\pi) - H(\rho) = K(\rho,\mathcal{U}) = \int \rho \ln \frac{\rho}{1/N} = \sum_i \rho_i \ln \rho_i + \ln N$$

$$= \ln N - H(\rho).$$