Basics about Entropy

Tianwei Gao

January 2025

1 Introduction

This is a note on basic properties and theorems about entropy in HDP.

Definition 1.1. (i)*Fix a random variable X, for a convex functionϕ we define H_ϕ as follows,*

$$H_\phi(x) := \mathbb{E}[\phi(X)] - \phi(\mathbb{E}X).$$

(ii) *When $\phi(u) = u \log u$, the entropy is defined by $H_\phi(e^X)$.*

Remark 1.1. (i) *For general convex function ϕ, $H_\phi(X)$ is called ϕ-entropy.*
(ii) *When $\phi(u) = u^2$, the ϕ-entropy gives variation.*

In the following, we keep the same notation as in HDS book. If not otherwise specified, the entropy function will be the ϕ-entropy with $\phi(u) = u \log u$, and we omit the subscript ϕ. We set $M_X(\lambda) := \mathbb{E}[e^{\lambda X}]$. A random variable X is said to satisfy the Bernstein entropy bound with $b, \sigma > 0$ if

$$H(e^{\lambda X}) \leq \lambda^2[-bM'_X(\lambda) + M_X(\lambda)(\sigma^2 - b\mathbb{E}X)], \quad \lambda \in [0, 1/b) \tag{1}$$

Here we present some basic properties of entropy function on parallel shifting, centering and rescaling.

Proposition 1.1. (i) *For a random variable X and a constant $C \in \mathbb{R}$,*

$$H(e^{\lambda(X+C)}) = e^{\lambda C} H(e^{\lambda X}).$$

(ii) *A random variable X, satisfies the Bernstein entropy bound for $b, \sigma > 0$ if and only if $\tilde{X} = X - \mathbb{E}[X]$ also satisfies the Bernstein entropy bound for b, σ.*
(iii)*For a zero-mean random variable X, X satisfies the Bernstein entropy bound with positive constants (b, σ) if and only if $\tilde{X} := \frac{X}{b}$ satisfies the Bernstein bound for $(\tilde{b}, \tilde{\sigma}) := (1, \frac{\sigma}{b})$*

Proof. We prove (ii). Note that $M_{\tilde{X}}(\lambda) = M_X(\lambda)e^{-\lambda\mathbb{E}X}$, then

$$M'_{\tilde{X}}(\lambda) = M'_X(\lambda)e^{-\lambda\mathbb{E}X} - \mathbb{E}X \cdot M_X(\lambda)e^{-\lambda\mathbb{E}X}$$

, substituting the above results into the Bernstein bound (1), we can see that the Bernstein bound for $\tilde{X}$ turns out to be

$$H(e^{\lambda\tilde{X}}) \leq e^{-\lambda\mathbb{E}X}\lambda^2[bM'_X(\lambda) + M_X(\lambda)(\sigma^2 - b\mathbb{E}X)],$$

By the formula of constant shifting (i), we can see that the above formula is equivalent to the Bernstein entropy bound for X. □

Now we present some explicit calculation of entropy.

Example 1.1 (Bounded Variable)**.** *Suppose X is a zero-mean bounded random variable supported on $[a, b]$, set $\sigma = b - a$, then*

$$H(e^{\lambda X}) \leq \frac{\lambda^2\sigma^2}{2} M_X(\lambda).$$

Proof. One can check that the following variational formulation of entropy holds,

$$H(e^{\lambda X}) = \inf_{t\in\mathbb{R}} \mathbb{E}[\psi(\lambda(X-t))e^{\lambda X}], \tag{2}$$

where $\psi(u) = e^{-u} - 1 + u$.

Then, note that for $u \geq 0$, we have

$$\phi(u) \leq \frac{u^2}{2},$$

take $t = a$, then

$$\psi(\lambda(X-t)) \leq \frac{\lambda^2(b-a)^2}{2},$$

we thus obtain the bound. ☐

Remark 1.2. *The constant can be sharpened to $\frac{1}{8}$, however we cannot take $t = \dfrac{a+b}{2}$ to achieve this($\psi(u)$ grows exponentially when $u < 0$), moreover, we did not use the zero-mean property in the above proof.*

Example 1.2 (Exponential family)**.** *Suppose random variable Y follows the exponential law*

$$p_\theta(y) = h(y)\exp(\langle\theta, T(y)\rangle - \Phi(\theta)),$$

assume that the regularization term $\Phi(\theta)$ is finite for all $\theta \in \mathbb{R}^n$, and its gradient $\nabla\Phi$ Lipschitz as a function of θ with Lipschitz constant L. Then for a vector of norm 1 v, the random variable

$$X := \langle v, T(y)\rangle$$

satisfies the following entropy bound:

$$H_\phi(e^{\lambda X}) \leq L\lambda^2 M_X(\lambda).$$

Proof. One can check that

$$M_X(\lambda) = e^{\Phi(\theta+\lambda v)-\Phi(\theta)},$$

then

$$\begin{aligned}
H(e^{\lambda X}) &= \lambda M_X'(\lambda) - M_X(\lambda)\log M_X(\lambda) \\
&= M_X(\lambda)(\lambda\nabla\Phi(\theta+\lambda v)\cdot v - (\Phi(\theta+\lambda v) - \Phi(\theta))) \\
&= M_X(\lambda)[\lambda(\nabla\Phi(\theta+\lambda v) - \nabla\Phi(\theta+\lambda\xi v))\cdot v] \\
&\leq L\lambda^2 M_X(\lambda)
\end{aligned}$$

where we used the intermediate value theorem and $\xi \in [0,1]$ is a non-negative constant. ☐

Proposition 1.2 (Variational formulation)**.** *Fix a random variable X, f a measurable function, $\lambda \in \mathbb{R}$, the entropy can be formed in the following variational viewpoint:*

$$H(e^{\lambda f(X)}) = \sup_{\mathbb{E}\exp(g(X))\leq 1} \mathbb{E}[g(X)e^{\lambda f(X)}].$$

Proof. When $g(X)$ is given by

$$g_0(X) := \lambda f(X) - \log\mathbb{E}[e^{\lambda f(X)}],$$

the equality holds. We now show that this is the optimal choice of g.

WLOG, in the following we assume $\lambda = 1$. Suppose g is a maxima, that is for any random variable h, $\mathbb{E}[e^{h(X)}] \leq 1$, and any positive real number ν, we have

$$\mathbb{E}[g(X)e^{\lambda f(X)}] > \mathbb{E}[\log(\frac{e^{g(X)} + \nu e^{h(X)}}{1+\nu})f(X)].$$

Let $\nu \to 0$, taking the derivative, we can see

$$\partial_\nu \mathbb{E}[\log(\frac{e^{g(X)} + \nu e^{h(X)}}{1+\nu})f(X)]|_{\nu=0} = \mathbb{E}[\frac{e^h}{e^g}e^f - e^f] \leq 0.$$

Take $h = f - \log \mathbb{E}[e^f]$, this becomes

$$\mathbb{E}[\frac{e^{2f}}{e^g}] \leq \mathbb{E}[e^f]^2.$$

Note that

$$\mathbb{E}[e^g] \leq 1$$
$$\mathbb{E}[e^{2f-g}] \geq \mathbb{E}[e^g]\mathbb{E}[e^{2f-g}] \geq \mathbb{E}[e^f]^2$$

By Cauchy Inequality, we have $e^g = \mathrm{Const} \cdot e^f$ almost surely, that is,

$$g - f = \mathrm{Const},$$

this shows that g_0 is optimal. □

Remark 1.3. (*i*) *In general, it is hard to find a similar variational formulation of entropy even with positive, monotone increasing, convex ϕ. Indeed, suppose the some entropy can be written in the following form, with a positive, monotone increasing, convex ϕ,*

$$\textit{Entropy} = \sup_{\mathbb{E}\phi(g) \leq \phi(0)} \mathbb{E}[g(X)\phi(f(X))]$$

To solve the optimized g would involve the inverse function of ϕ', and the optimized g is required to satisfy the following relation,

$$\phi'(g(X)) \propto \phi(f(X)),$$

the regularization constant (corresponding to the term $\log \mathbb{E}[e^{\lambda f}]$ in our case) is in general not tractable.

Lemma 1.3 (Tensorization of Entropy)**.** *Assume $X_1, \ldots, X_n$ are independent random variables, $f : \mathbb{R}^n \to \mathbb{R}$, then*

$$H(e^{\lambda f(X_1,\ldots,X_n)}) \leq \sum_{k=1}^{n} \mathbb{E}[H(e^{\lambda f_k(x_k)})|X_{\backslash k}]$$

Remark 1.4. *With the notation same as in variational formulation, for $g(X)$ such that $\mathbb{E}[e^{g(X)}] \leq 1$, we can define*

$$g_k(X) = \log \frac{\mathbb{E}[e^{g(X)}|X_k, \ldots, X_n]}{\mathbb{E}[e^{g(X)}|X_{k+1}, \ldots, X_n]},$$

then $\mathbb{E}[\exp(g_k(X_k, \ldots, X_n))|X_{k+1}, \ldots, X_n] = 0$, and

$$g(X) \leq \sum_{k=1}^{n} g_k(X).$$

Now the following procedure is very similar to the proof of tensorization of variation.