熵方法与 Herbst 论证

MIStatlE

更新：2025 年 4 月 22 日

1 引言

在高维概率与机器学习中，我们经常需要控制

$$Z = f(X_1, \dots, X_n)$$

偏离其期望 $\mathbb{E}[Z]$ 的概率。经典工具如 Hoeffding、Bernstein 和 McDiarmid 不等式，本质上依赖对数矩母函数（MGF）的张量化，但一般非线性函数 f 的 MGF 难以拆分。熵方法的核心在于用熵替代 MGF，以利用熵的可加性进行坐标级别的控制。

深入理解

熵方法的核心是将对 MGF 的控制转换为对熵 $\mathrm{Ent}[\cdot]$ 的控制，因为后者具有天然的张量化性质：

$$\mathrm{Ent}\big[g(X_1, \dots, X_n)\big] \le \sum_{k=1}^{n} \mathbb{E}\big[\mathrm{Ent}_k\big(g(X_1, \dots, X_n)\big)\big],$$

每一项只涉及第 k 坐标的随机性。

2 熵的定义与基本性质

2.1 形式定义

定义（熵）

令 $Z > 0$ 为正随机变量，定义

$$\mathrm{Ent}[Z] = \mathbb{E}\big[Z \log Z\big] - \mathbb{E}[Z] \log \mathbb{E}[Z].$$

备注 2.1. 令 $\phi(t) = t \log t$，则

$$\mathrm{Ent}[Z] = \mathbb{E}[\phi(Z)] - \phi(\mathbb{E}[Z]) = \mathbb{E}\big[D(Z \| \mathbb{E}[Z])\big],$$

其中 $D(y \| x) = \phi(y) - \phi(x) - \phi'(x)(y - x)$ 是 *Bregman* 距离，由于 $\phi(t) = t \log t$ 在 $(0, \infty)$ 上凸，从而 $\mathrm{Ent}[Z] \ge 0$。

2.2 熵与对数矩母函数的联系

定义对数矩母函数

$$\psi(\lambda) = \log \mathbb{E}\big[e^{\lambda(X - \mathbb{E}[X])}\big].$$

引理（熵–MGF 公式）

对于任意 $\lambda \in \mathbb{R}$，有

$$\frac{\operatorname{Ent}\left[e^{\lambda X}\right]}{\mathbb{E}[e^{\lambda X}]} = \lambda\,\psi'(\lambda) - \psi(\lambda).$$

证明. 设 $M(\lambda) = \mathbb{E}[e^{\lambda X}]$，则

$$\operatorname{Ent}[e^{\lambda X}] = \mathbb{E}[e^{\lambda X}\,\lambda X] - M(\lambda)\log M(\lambda) = \lambda M'(\lambda) - M(\lambda)\log M(\lambda).$$

两边除以 $M(\lambda)$ 并记 $\psi(\lambda) = \log M(\lambda)$，即得

$$\frac{\operatorname{Ent}[e^{\lambda X}]}{M(\lambda)} = \lambda\,\frac{M'(\lambda)}{M(\lambda)} - \psi(\lambda) = \lambda\,\psi'(\lambda) - \psi(\lambda).$$

□

示例 2.1. 若 $X \sim \mathcal{N}(0, \sigma^2)$，则

$$\operatorname{Ent}[e^{\lambda X}] = \frac{\lambda^2\sigma^2}{2}\,\mathbb{E}[e^{\lambda X}].$$

2.3 熵的变分公式与交换不等式

引理（熵的变分公式）

对于任意正随机变量 Z，

$$\operatorname{Ent}[Z] = \sup_{\substack{H:\\ \mathbb{E}[e^H]=1}} \mathbb{E}[Z\,H].$$

证明. 令 $H = \log(Z/\mathbb{E}[Z])$，则 $e^H = Z/\mathbb{E}[Z]$ 且 $\mathbb{E}[e^H] = 1$，此时

$$\mathbb{E}[Z\,H] = \mathbb{E}[Z\log(Z)] - \mathbb{E}[Z]\log\mathbb{E}[Z] = \operatorname{Ent}[Z].$$

对于任意满足 $\mathbb{E}[e^H] = 1$ 的 H，由 Jensen 或 KL 散度非负性可得 $\mathbb{E}[Z\,H] \le \operatorname{Ent}[Z]$，故上确界即为 $\operatorname{Ent}[Z]$。 □

引理（交换不等式）

若正随机变量 Z, Z' 独立同分布，则

$$\operatorname{Ent}[Z] \le \frac{1}{2}\,\mathbb{E}\big[(Z - Z')(\log Z - \log Z')\big].$$

证明. 由变分公式，

$$\operatorname{Ent}[Z] = \sup_{\mathbb{E}[e^H]=1} \mathbb{E}[Z\,H] \ge \mathbb{E}\Big[Z \cdot \tfrac{1}{2}(\log Z - \log Z')\Big],$$

其中取 $H = \frac{1}{2}(\log Z - \log Z')$，可校正常数保证 $\mathbb{E}[e^H] = 1$。对称地，

$$\operatorname{Ent}[Z'] \ge \mathbb{E}\Big[Z' \cdot \tfrac{1}{2}(\log Z' - \log Z)\Big].$$

将两式相加并因对称性可得所述不等式。 □

3 Herbst 方法：熵到次高斯尾界

定理（Herbst 不等式）

若存在常数 $\nu > 0$ 使得对所有 $\lambda \geq 0$，

$$\mathrm{Ent}[e^{\lambda X}] \leq \frac{\lambda^2 \nu^2}{2}\,\mathbb{E}[e^{\lambda X}],$$

则

$$\psi(\lambda) = \log \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \leq \frac{\lambda^2\nu^2}{2},$$

即 $X - \mathbb{E}X$ 为参数 ν 的次高斯随机变量。

证明. 由熵–MGF 公式可得

$$\frac{d}{d\lambda}\Big(\frac{\psi(\lambda)}{\lambda}\Big) = \frac{\lambda\psi'(\lambda) - \psi(\lambda)}{\lambda^2} = \frac{\mathrm{Ent}[e^{\lambda X}]}{\lambda^2\,\mathbb{E}[e^{\lambda X}]} \leq \frac{\nu^2}{2}.$$

又 $\lim_{\lambda\to 0}\psi(\lambda)/\lambda = 0$，故

$$\frac{\psi(\lambda)}{\lambda} \leq \int_0^\lambda \frac{\nu^2}{2}\,d\xi = \frac{\lambda\nu^2}{2},$$

乘以 λ 即得 $\psi(\lambda) \leq \frac{\lambda^2\nu^2}{2}$。 □

备注 3.1. 由于 $\mathrm{Ent}[e^{\lambda(X+c)}] = e^{\lambda c}\,\mathrm{Ent}[e^{\lambda X}]$，*Herbst* 条件对平移不变，无需中心化。

4 熵的张量化

定理（熵的张量化）

设 $X_1, \dots, X_n$ 独立，$g : \mathcal{X}^n \to [0,\infty)$，定义

$$\mathrm{Ent}_k[g] = \mathrm{Ent}\big(g(X_1,\dots,X_{k-1},X_k,X_{k+1},\dots,X_n)\big).$$

则

$$\mathrm{Ent}[g(X_1,\dots,X_n)] \leq \sum_{k=1}^n \mathbb{E}[\mathrm{Ent}_k[g]].$$

证明. 令 $Z = g(X_1,\dots,X_n)$，并定义

$$U_k = \log\mathbb{E}[Z \mid X_1,\dots,X_k] - \log\mathbb{E}[Z \mid X_1,\dots,X_{k-1}].$$

由全概率展开和变分公式，

$$\mathrm{Ent}[Z] = \mathbb{E}\big[Z(\log Z - \log\mathbb{E}Z)\big] = \sum_{k=1}^n \mathbb{E}[Z\,U_k].$$

固定其余坐标，$Z\,U_k$ 视作仅关于 X_k 的函数，由变分公式得 $\mathbb{E}[Z\,U_k] \leq \mathbb{E}[\mathrm{Ent}_k[Z]]$，求和即得结论。 □

深入理解

张量化定理展示了熵的“可加性”，使我们能逐坐标控制复杂函数的整体波动。

5 应用：推广 McDiarmid 不等式

命题（熵方法版 McDiarmid）

若 $f:\mathcal{X}^n\to\mathbb{R}$ 满足对每个 k，

$$|f(\dots,x_k,\dots)-f(\dots,x_k',\dots)|\le L_k,$$

则对任意 $t\ge 0$，

$$\mathbb{P}\big(|f(X)-\mathbb{E}[f(X)]|\ge t\big)\le 2\exp\Big(-\frac{2t^2}{\sum_{k=1}^n L_k^2}\Big).$$

证明. 令 $g=e^{\lambda f(X)}$。由单坐标修正 log–Sobolev（或 Hoeffding 引理），

$$\mathrm{Ent}_k[g]\le\frac{\lambda^2L_k^2}{8}\mathbb{E}[g].$$

张量化得

$$\mathrm{Ent}[g]\le\frac{\lambda^2}{8}\sum_{k=1}^n L_k^2\mathbb{E}[g].$$

由 Herbst 定理 $f-\mathbb{E}f$ 为次高斯，参数 $\nu^2=\frac{1}{4}\sum L_k^2$。再用 Chernoff 法，

$$\mathbb{P}(f-\mathbb{E}f\ge t)\le\exp\big(-t^2/(2\nu^2)\big)=\exp\big(-2t^2/\sum L_k^2\big),$$

对称地乘以 2 即得结果。 □

示例 5.1. 令 $\varepsilon_k\in\{\pm1\}$ *i.i.d.*，$Z=\sup_{a\in A}\sum a_k\varepsilon_k$，则

$$|Z(\varepsilon)-Z(\varepsilon')|\le\sup_{a\in A}|a_k|,$$

直接套用命题，并可优化常数常数。

6 小结

熵方法通过“定义 →Herbst 论证 → 张量化”三步，为非线性函数提供了统一的浓度不等式框架。后续可结合 log–Sobolev、Talagrand 输运等进一步深化。