A Unified View on Learning Unnormalized Distributions via Noise-Contrastive Estimation

Jongha (Jon) Ryu
Abhin Shah
Gregory W. Wornell
MIT EECS | correspondence to: jongha@mit.edu

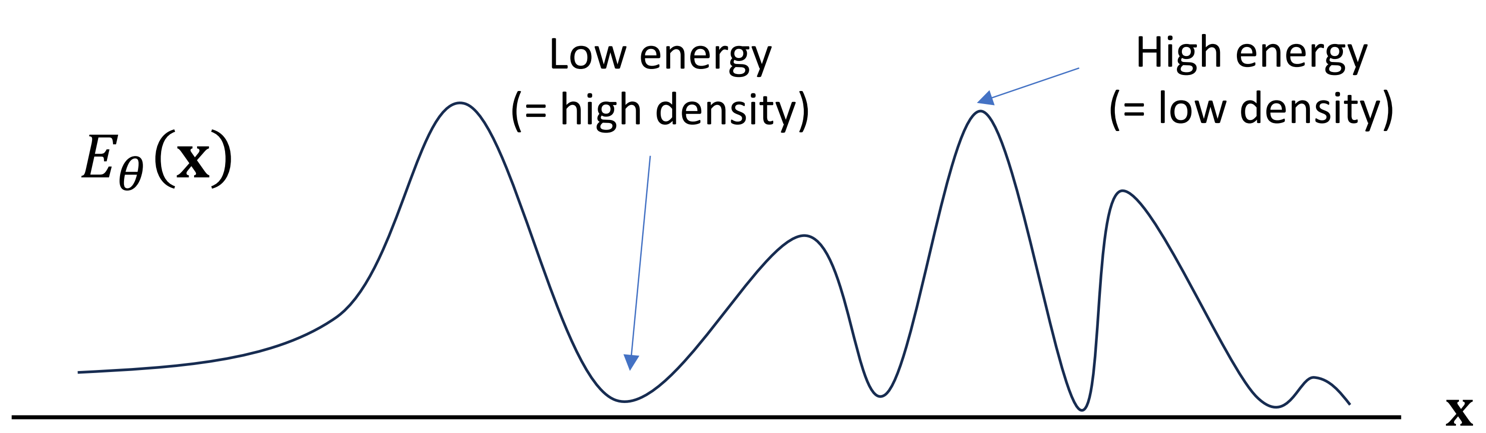
TL;DR: We unify various methods for learning EBMs via a family of NCE principles and establish finite-sample rates for exponential family distributions

cf. noise contrastive estimation, score matching, interaction screening objectives, pseudo maximum likelihood

Unnormalized Distributions (a.k.a. Energy-Based Models (EBMs)) Are Flexible Probabilistic Models

$$p_\theta(\mathbf{x}) \propto \exp(-\underbrace{E_\theta(\mathbf{x})}_{\text{“energy”}})$$

canonical example: exponential family $E_\theta(\mathbf{x}) = -\theta^\intercal \psi(\mathbf{x})$



Application scenarios: graphical models (Markov random fields), physical modeling (Ising models), causal modeling, image modeling, …

Challenge: Unknown Normalization

Flexibility comes at the cost of not knowing the normalization constant (a.k.a. partition function)

$$Z_\theta \stackrel{\text{def}}{=} \int \exp(-E_\theta(\mathbf{x}))d\mathbf{x} \qquad p_\theta(\mathbf{x}) = \frac{\exp(-E_\theta(\mathbf{x}))}{Z_\theta}$$

$\Rightarrow$ **Can't apply maximum likelihood estimation** (MLE)!

Some known approaches

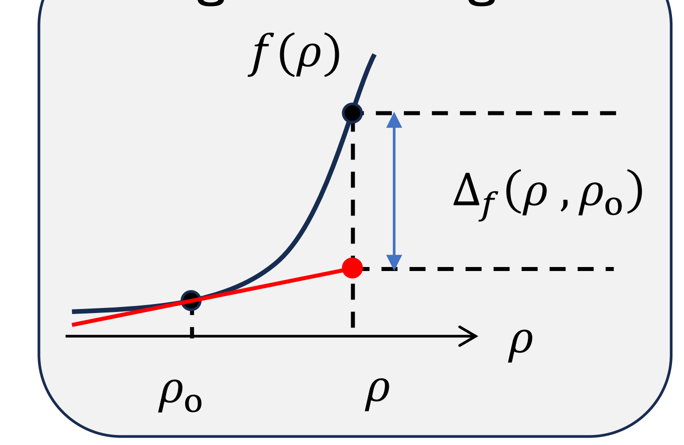
- Noise-contrastive estimation (Gutmann & Hyvarinen, 2010)
- Score matching (Hyvarinen, 2012)
- Contrastive divergence (Hinton, 2002)
- Variants and other proposals in some specific settings (pseudo likelihood (Besag, 1975), interaction screening (Vuffray et al., 2016), …)

Q. Is there a unifying principle?

Noise Contrastive Estimation (NCE) Is A Unifying Principle!

- Unnormalized model $\phi_\theta(\mathbf{x}) \stackrel{\text{def}}{=} \exp(-E_\theta(\mathbf{x}))$
- Data distribution $q_{\mathrm{d}}(\mathbf{x})$
- Noise distribution $q_{\mathrm{n}}(\mathbf{x})$ (typically Gaussian)
- **Idea**: match **density ratios** $\dfrac{q_{\mathrm{d}}(\mathbf{x})}{q_{\mathrm{n}}(\mathbf{x})} \approx \dfrac{c\phi_\theta(\mathbf{x})}{q_{\mathrm{n}}(\mathbf{x})}$

Bregman divergence



For a strictly *convex* function f and $\nu > 0$, (Pihlaja et al., 2010)

$$(f\text{-NCE obj.}) \quad \mathcal{L}_f^{\mathrm{nce}}(\phi_\theta, c) \stackrel{\text{def}}{=} \mathbb{E}_{q_{\mathrm{n}}(\mathbf{x})}\left[\Delta_f\left(\frac{q_{\mathrm{d}}(\mathbf{x})}{\nu q_{\mathrm{n}}(\mathbf{x})}, \frac{c\phi_\theta(\mathbf{x})}{\nu q_{\mathrm{n}}(\mathbf{x})}\right)\right] - (\text{const.})$$

Note: Can be unbiasedly estimated by samples from $q_{\mathrm{d}}(\mathbf{x})$ and $q_{\mathrm{n}}(\mathbf{x})$
Caveat: Need to choose $q_{\mathrm{n}}(\mathbf{x})$ *carefully*

We propose and study **TWO variants**

Variant 1. α-Centered NCE

- Consider $f_\alpha(\rho) = \frac{\rho^\alpha - 1}{\alpha(\alpha-1)}$ ($\alpha \notin \{0,1\}$) (“asymmetric power”)
- Define a **normalized** model (which we call “α-centered model”)

$$\tilde{\phi}_{\theta;\alpha}(\mathbf{x}) \stackrel{\text{def}}{=} \frac{\phi_\theta(\mathbf{x})}{Z_\alpha(\theta)}, \qquad \text{where } Z_\alpha(\theta) \stackrel{\text{def}}{=} \left(\mathbb{E}_{q_{\mathrm{n}}(\mathbf{x})}\left[\left(\frac{\phi_\theta(\mathbf{x})}{q_{\mathrm{n}}(\mathbf{x})}\right)^\alpha\right]\right)^{1/\alpha}$$

such that it's α-centered: $\mathbb{E}_{q_{\mathrm{n}}(\mathbf{x})}\left[\left(\frac{\tilde{\phi}_{\theta;\alpha}(\mathbf{x})}{q_{\mathrm{n}}(\mathbf{x})}\right)^\alpha\right] = 1$

$$(\alpha\text{-CentNCE obj.}) \quad \mathcal{L}_\alpha^{\mathrm{cent}}(\phi_\theta) \stackrel{\text{def}}{=} \mathcal{L}_{f_\alpha}^{\mathrm{nce}}(\tilde{\phi}_{\theta;\alpha}, 1)$$

Objectives	$\alpha = 0$	$\alpha = \frac{1}{2}$	$\alpha = 1$
f_α-NCE	$\mathbb{E}_{q_{\mathrm{d}}}[\frac{q_{\mathrm{n}}}{\phi_\theta}] + \mathbb{E}_{q_{\mathrm{n}}}[\log \frac{\phi_\theta}{q_{\mathrm{n}}}]$ (InvIS (Pihlaja et al., 2010))	$2(\mathbb{E}_{q_{\mathrm{d}}}[\sqrt{\frac{q_{\mathrm{n}}}{\phi_\theta}}] + \mathbb{E}_{q_{\mathrm{n}}}[\sqrt{\frac{\phi_\theta}{q_{\mathrm{n}}}}])$ (eNCE (Liu et al., 2021))	$\mathbb{E}_{q_{\mathrm{d}}}[\log \frac{q_{\mathrm{n}}}{\phi_\theta}] + \mathbb{E}_{q_{\mathrm{n}}}[\frac{\phi_\theta}{q_{\mathrm{n}}}]$ (Importance Sampling (IS) (Pihlaja et al., 2010; Riou-Durand & Chopin, 2018))
α-CentNCE	$\mathbb{E}_{q_{\mathrm{d}}}[\frac{q_{\mathrm{n}}}{\phi_\theta}]e^{\mathbb{E}_{q_{\mathrm{n}}}[\log \frac{\phi_\theta}{q_{\mathrm{n}}}]}$ (**GlobalGISO** (Shah et al., 2023))	$2\mathbb{E}_{q_{\mathrm{d}}}[\sqrt{\frac{q_{\mathrm{n}}}{\phi_\theta}}]\mathbb{E}_{q_{\mathrm{n}}}[\sqrt{\frac{\phi_\theta}{q_{\mathrm{n}}}}]$	$\mathbb{E}_{q_{\mathrm{d}}}[\log \frac{q_{\mathrm{n}}}{\phi_\theta}] + \log \mathbb{E}_{q_{\mathrm{n}}}[\frac{\phi_\theta}{q_{\mathrm{n}}}]$ (**MLE** (Fisher, 1922), **MC-MLE** (Geyer, 1994; Jiang et al., 2023))

Takeaway 1
α-CentNCE unifies MLE, MC-MLE (Geyer, 1994), and GlobalGISO (Shah et al., 2023) as limiting instances

Variant 2. f-Conditional NCE

- $q_{\mathrm{n}}(\mathbf{x})$, so “conditional NCE” (Ceylan and Gutmann, 2018) was proposed
- Conditional distribution $\pi(\mathbf{y}|\mathbf{x})$ (ex: $\mathbf{y} = \mathbf{x} + \epsilon\mathbf{v}$, $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathrm{I})$)
- **Idea**: match **“joint” density ratios** $\dfrac{q_{\mathrm{d}}(\mathbf{x})\pi(\mathbf{y}|\mathbf{x})}{q_{\mathrm{d}}(\mathbf{y})\pi(\mathbf{x}|\mathbf{y})} \approx \dfrac{\phi_\theta(\mathbf{x})\pi(\mathbf{y}|\mathbf{x})}{\phi_\theta(\mathbf{y})\pi(\mathbf{x}|\mathbf{y})}$
- We extend CondNCE to f-CondNCE in the same way as from NCE to f-NCE
- For Gaussian $\pi(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{x}, \epsilon^2\mathrm{I})$), Ceylan and Gutmann (2018) argued

$$\lim_{\epsilon \to 0}(\text{CondNCE}) = (\text{Score Matching})$$

- **Takeaway 2**: However, we reveal that, for any convex f:
  (1) This is only true in the population limit;
  (2) With finite samples, **f-CondNCE is dominated by statistical noise** as $\epsilon \to 0$
  (so NOT equivalent to SM)

Takeaway 3. Finite-Sample Analysis

- For exponential family $\phi_\theta(\mathbf{x}) = \exp(\theta^\intercal \psi(\mathbf{x}))$, we analyze the finite-sample performance of the estimators f-NCE, α-CentNCE, f-CondNCE
- Most are **first** convergence rate guarantees for the considered estimators
- This is thanks to our unifying framework
- **Idea:** Adapt the convergence rate analysis of GlobalGISO (Shah et al., 2023)

- **Remark**: We can also extend this unification to *local* models (e.g., sparse Markov random fields), which unifies pseudo likelihood (Besag, 1975), and ISO (Vuffray et al., 2016; 2021; Shah et al., 2021a; Ren et al., 2021)