On Fenchel Mini-Max Learning

Chenyang Tao, Duke University

Oct 27th, 2019

Background

Maximal likelihood estimation (MLE)

- MLE is perhaps the most important workhorse in statistical inference and machine learning
  - It identifies the most probable model $\hat{\theta}_{\text{MLE}}$ via maximizing the expected of model $\log$-likelihood wrt empirical observations

$$L(\theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} \log p(x_i; \theta). \tag{1}$$

  - Which in term minimizes the KL-divergence $\text{KL}(\hat{p}_d \parallel p_\theta)$ between empirical distribution $\hat{p}_d$ and model distribution p_θ

Background

Challenges of MLE with non-normalized statistical models

- In many cases, statistical models are given in the form of non-normalized exponential family $\tilde{p}(x;\theta) = \exp(-\psi(x;\theta))$
  - $\psi(x;\theta)$ is known as the *potential function*.
  - PDF is known up to a multiplicative constant $Z(\theta)$,

$$p(x;\theta) = \frac{1}{Z(\theta)}\tilde{p}(x;\theta), \tag{2}$$

  - $Z(\theta) = \int \tilde{p}(x';\theta)\,\mathrm{d}x'$ is called the *partition function*
  - In general, $Z(\theta)$ is **analytically intractable**

Background

Likelihood-based approaches for generative modeling

- Generative modeling tries to learn a stochastic procedure that synthesizes realistic samples
- Despite the huge success of adversarial solutions (GANs), likelihood-based models remain a popular choice
  - stability, model interpretability, etc.
- Broadly categorized into explicit and approximate approaches
  - Explicit: normalizing flows
    - invertible transformations with tractable Jacobian, **costly**
  - Approximate: variational inference (VI)
    - evidence lower bound of likelihood, **loose**

Background

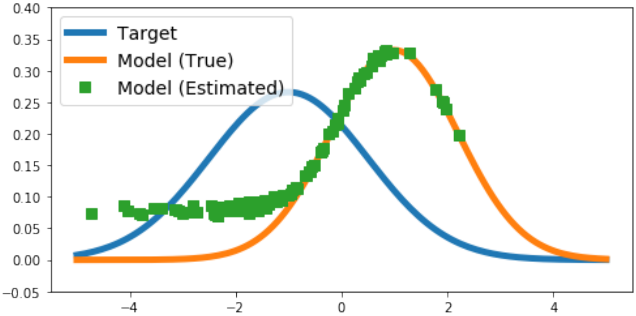
Existing solutions for MLE of non-normalized models

- MCMC-based
  - MCMC-MLE (Geyer 1991), contrastive divergence (Hinton 2002)
- Laplacian-based
  - Score matching (Hyvarinen 2005)
- Discriminative
  - Noise contrastive estimation (Gutmann 2010), generalized Bregman estimation (Gutmann 2012)
- Others
  - Stein (not exactly), dynamic embedding (Dai 2018)

Mini-Max Likelihood Estimation (Ours)

General idea

- Inspired by adversarial training, we exploited the Fenchel conjugacy and reformulate MLE into a mini-max game
  - *Min-game* recovers the model likelihood
  - *Max-game* matches model to the data distribution



Mini-Max Likelihood Estimation (Ours)

Our MMLE model

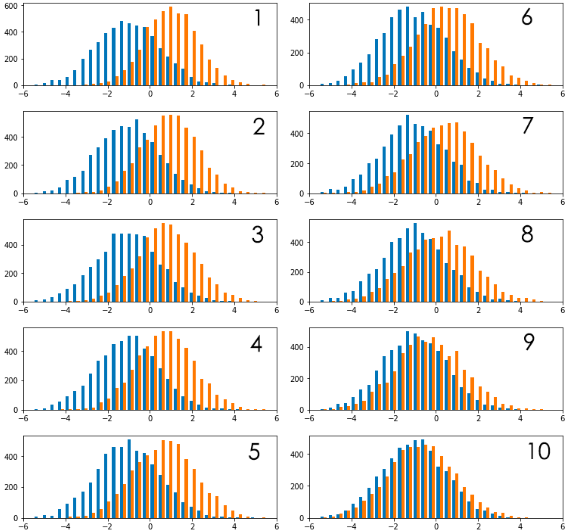
$$\hat{\psi}_{\text{MLE}} = \arg\max_{\psi} \left\{ -\min_{\boldsymbol{u}} \left\{ \sum_{i=1}^{n} \left(u_i + e^{-u_i} I(x_i; \psi) \right) \right\} \right\} \tag{3}$$

- Non-normalized model distribution $\tilde{p}_{\psi}(x) = \exp(-\psi(x))$
- Importance-weighted estimator $I(x; \psi)$

$$I(x; \psi_{\theta}) = \int \left(\frac{1}{q(x')} e^{\psi(x) - \psi(x')} \right) q(x') \, \mathrm{d}x' \tag{4}$$

  - $q(x)$ is the proposal distribution
- Auxiliary variable u_i for each data point x_i and Min-game returns the normalized $\log$-likelihood, *e.g.* $u^* = \log p_{\psi}(x)$

Demo with Simple Gaussian



Application to Generative Models

Applying the same trick to latent variable model

$$\arg\max_{\alpha,\beta} \left\{ \min_{\boldsymbol{u}} \left\{ \sum_{i=1}^{n} \left(u_i + e^{-u_i} \int q_\beta(z|x) \left\{ \frac{p_\alpha^{\tau_t}(x,z)}{q_\beta(z|x)} \right\} \mathrm{d}z \right) \right\} \right\} \quad (5)$$

- $q_\beta(z|x)$ resembles the approximate posterior in VI
  - In IW-VAE it also functions as proposal distribution
- τ_t is the annealing factor
- *In contrast to VI, we optimize an estimate of the* $\log$*-likelihood rather than a lower bound*

Going Forward

- Amortizing auxiliary variable with a deep neural net $u(x; \phi)$
- Using $u(x; \phi)$ as a critic in adversarial distribution matching
- Introducing inductive bias to improve generalization
- Reasonably confident to have a draft ready with preliminary quantitative results (on real data) by next weekend