On Fenchel Mini-Max Learning

Chenyang Tao[1], Liqun Chen[1], Shuyang Dai[1], Junya Chen[1,2], Ke Bai[1], Dong Wang[1], Jianfeng Feng[3], Wenlian Lu[1], Georgiy Bobashev[4], Lawrence Carin[1]
{chenyang.tao, lcarin}@duke.edu

[1]Electrical & Computer Engineering, Duke University [2]School of Mathematical Sciences, Fudan University [3]ISTBI, Fudan University [4]RTI International

Dec 10, 2019 @ NeurIPS, Vancouver, BC, Canada

Outline

1. Background
2. Non-Parametric Maximal Likelihood Estimation
3. Fenchel Mini-Max Learning (Mini-Max MLE)
4. Experiments

Goals of Probabilistic Modeling

Four desiderata of probabilistic modeling

- *Estimation* ✓ $\Leftarrow$ Traditional focuses
  - Find a parameterized model $\mathcal{M}_\theta$ that best describes the data
- *Inference* ✓
  - Identify the latent properties (e.g. group identity) of data
- *Sampling* ✓ $\Leftarrow$ **Recent interests**
  - Emulating a stochastic rule that generates data
- *Likelihood evaluation (evidence)* ✗
  - Evaluate the likelihood for a given sample (*e.g.*, outlier)
- *Other concerns*
  - model scalability, training stability, expressiveness, *etc.*

✓ : frequently used ✗: less frequently used

Goals of Probabilistic Modeling

Four desiderata of probabilistic modeling

- *Estimation* ✓
- *Inference* ✓
- *Sampling* ✓
- *Likelihood evaluation (evidence)* ✗
- *Other concerns*
  - model scalability, training stability, expressiveness, *etc.*

✓: frequently used ✗: less frequently used

- **We want to present a scheme without any major compromise on the above points.**

Our Contributions

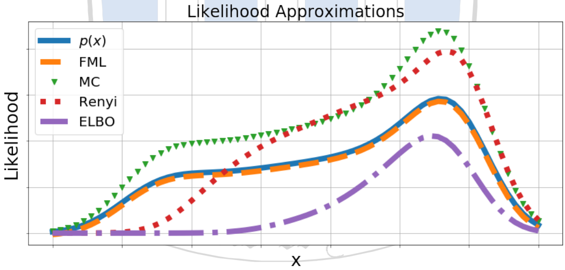
We present Fenchel Mini-Max Learning (FML) which highlights

1. A **Mini-Max formulation** of *Maximal Likelihood Estimation*
2. **Unbiased** likelihood estimator directly amendable to SGD
3. Amortized estimation with deep neural networks
4. A latent-variable model competitive to variational inference

- We show FML compare favorably to existing alternatives in likelihood-based distribution learning across a wide range of applications, such as
  - density estimation, generative modeling, natural language processing, reinforcement learning, etc.

A Comparison to Competing Likelihood Estimators

- Naïve MC over-estimates: $\log \hat{p}_{\theta,\text{MC}}(x) \geq \log p_\theta(x)$
- Evidence lower bound under-estimates: $\text{ELBO}_\theta(x) \leq \log p_\theta(x)$
- Rényi bound approximates: $\text{RVI}_\theta(x) \approx \log p_\theta(x)$
- FML estimate is tight: $\text{FML}_\theta(x) = \log p_\theta(x)$



Maximal Likelihood Criteria

Maximum likelihood estimation (MLE)

- $\{x_i\}_{i=1}^n$ are independent observations of true distribution $p_d(x)$
- $\{p_\theta(x)\}_{\theta\in\Theta}$ is a family of distributions parameterized by α
- $\log$-likelihood loss: $\hat{\ell}(\theta) = \sum_i \log p_\theta(x_i)$
- Maximizing expected likelihood:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta\in\mathcal{A}}\{\ell(\theta)\} \tag{1}$$

- This is equivalent to minimizing $\text{KL}(p_d \parallel p_\theta)$

Likelihood-based Probabilistic Modeling

One-step approach

- Direct construction of a parameterized stochastic procedure with tractable likelihoods. (*e.g.*, generative flows)

Two-step approach

- First estimate an (unnormalized) density with empirical samples (*e.g.*, non-parametric density estimation)
- To draw new samples, use established sampling procedure (*e.g.*, MCMC) to sample from the density estimate
- **This study addresses the challenge of estimation an unnormalized density**

Background on Estimating unnormalized Statistical Models

Challenges of MLE with unnormalized statistical models

- In many cases, statistical models are given in the form of unnormalized 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**

Density Estimation of Unnormalized Statistical Models

The challenge of density estimation for unnormalized models

- In MLE training, one optimizes
$$\log p_\theta(x) = -\psi_\theta(x) - \log Z_\theta \tag{3}$$
  - Naïve solution: directly estimate Z_θ
- Finite sample MC estimate bias the likelihood estimate $\log \hat{p}_\psi(x)$ because the integral is within the $\log$
$$\log \hat{p}_\theta(x) = -\psi_\theta(x) - \log \hat{Z}_\theta, \tag{4}$$
$$\log \hat{Z}_\theta = \log\left(\frac{1}{m}\sum_{j=1}^{m} \exp(-\psi_\theta(X_j'))\right), \tag{5}$$
$$\mathbb{E}_{X_j}[\log \hat{Z}_\theta] \leq \log(\mathbb{E}_{X_j}[\hat{Z}_\theta]) = \log Z_\theta, \text{(Jensen's ineq)} \tag{6}$$
$$\mathbb{E}_{X_j}[\log \hat{p}_\theta(x)] \geq \log p_\theta(x). \tag{7}$$

Density Estimation of Unnormalized Statistical Models

Existing solutions

- Potential-based
  - MCMC-MLE [Geyer, 1991]
  - Contrastive divergence (CD), [Hinton, 2002]
  - Noise Contrastive Estimation (NCE), [Gutmann, 2010]
  - Dynamic dual embedding (DDE), [Dai, 2018]
- Score-based
  - Score matching (SM), [Hyvarinen, 2005]
  - De-noising auto-encoder (DAE), [Alain, 2014]
  - Stein implicit learning (SIL), [Li, 2018]
- **Trade-offs are made in terms of the goals one wants to achieve, because $\cdots$**

NO single method hits all bullets : - (

Table: Comparison of popular probabilistic modeling procedures.

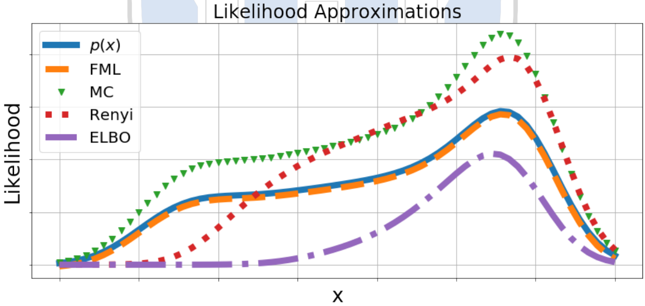
Model	Inference	Sampling	Likelihood	Scalability
MCMC	No	Yes	No	Poor
RBM	Yes	Yes	No	Good
SM	No	No	No	Poor
NCE	No	No	Estimate	Depends
DDE	No	Yes	Exact	Medium
VI	Yes	Yes	Bound	Good
FLOW	No	Yes	Exact	Tricky
SVGD	No	Yes	No	Medium
GAN	No	Yes	No	Good
OT	No	Yes	No	Depends
FML	**Yes**	**Yes**	**Estimate**	**Good**

Fenchel 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

(See our paper for details on how all four goals are accommodated)



Fenchel Mini-Max Learning

Fenchel Conjugacy

- Let $f(t)$ be a proper convex, lower-semicontinuous function
- The convex conjugate function $f^*(v)$ is defined as

$$f^*(v) = \sup_{t \in \mathcal{D}(f)} \{tv - f(t)\} \tag{8}$$

  - $\mathcal{D}(f)$ denotes the domain of function f
  - f^* is again convex and lower-semicontinuous.
  - Fenchel conj. pair (f, f^*) are dual to each other $((f^*)^* = f)$
- Example: $(-\log(t), -1 - \log(-v))$ is a Fenchel pair
- Change of variable $v = \exp(-u)$ gives

$$-\log(t) = \max_u \{-u - \exp(-u)t + 1\} \tag{9}$$

Mini-Max Likelihood Estimation

Fenchel Mini-Max Formulation

- For unnormalized model distribution $\tilde{p}_\psi(x) = \exp(-\psi(x))$

$$\hat{\psi}_{\text{MLE}} = \arg\max_{\psi} \left\{ -\min_{\boldsymbol{u}} \left\{ \sum_{i=1}^{n} (u_i + e^{-u_i} I(x_i; \psi)) \right\} \right\} \tag{10}$$

- Importance-weighted estimator $I(x; \psi)$

$$I(x; \psi_\theta) = \int \left(\frac{1}{q(x')} e^{\psi(x) - \psi(x')} \right) q(x') \, \mathrm{d}x' \tag{11}$$

  - $q(x)$ is the proposal distribution

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

Understanding Fenchel Mini-Max Likelihood Estimation

A gradient analysis

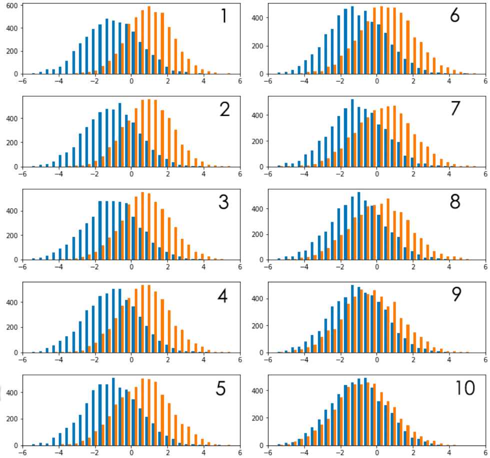
- In standard MLE learning, we have $\nabla \log p_\theta(x) = \frac{\nabla p_\theta(x)}{p_\theta(x)}$
  - The gradient of likelihood is normalized by model evidence
- Insight: while $\nabla p_\theta(x)$ is difficult to compute (because of Z_θ), unbiased gradient estimate of the inverse likelihood $\frac{1}{p_\theta(x)}$ is easy to get $\nabla \left\{ \frac{1}{p_\theta(x)} \right\} = \int \nabla \{\exp(\psi_\theta(x) - \psi_\theta(x'))\} \, \mathrm{d}x'$
- This connects FML gradient via

$$\begin{aligned} \nabla J_\theta(x; \hat{u}_x, \psi) &= -\nabla \left\{ \exp(-\hat{u}_x) \int e^{\psi_\theta(x) - \psi_\theta(x')} \, \mathrm{d}x' \right\} \\ &= -\hat{p}_\theta(x) \nabla \left\{ \frac{1}{p_\theta(x)} \right\} = \frac{\hat{p}_\theta(x)}{p_\theta(x)} \nabla \log p_\theta(x) \approx \nabla \log p_\theta(x), \end{aligned} \tag{12}$$

- It's easy to show when the likelihood ratio is bounded, we can recover the same exact MLE solution

Learning with FML: Simple Gaussian

(blue: data, orange: model)



Question: Does FML converge to ground-truth?

Convergence Guarantees of FML

Proposition (Convergence of FML with stochastic gradient descent)

- Let $\{\eta_t\}$ be the scheduled learning rate and $\{\xi_t\}$ be the approximation error. If the generalized learning rate $\tilde{\eta}_t = \eta_t \xi_t$ satisfies $\sum_t \mathbb{E}[\tilde{\eta}_t] = \infty$ and $\sum_t \mathbb{E}[\tilde{\eta}_t^2]$, we have
  - *(Convex setting)* if the likelihood loss has a unique equilibrium point and is asymptotically stable (e.g. strictly convex), then under standard Robbins-Monro regularity conditions FML-SGD converges to MLE solution $\hat{\theta}_{\text{MLE}}$ with probability 1 from any θ_0.
  - *(Non-convex setting)* if the likelihood loss has Lipschitz-continuous, then FML-SGD will converge to a stationary point of $f(\theta)$ with probability 1, *i.e.*, $\|\nabla_\theta f(\theta)\| \to 0$ as $t \to 0$.

Fenchel Mini-Max Learning for Latent Variable Models

Applying the same trick to latent variable model

$$\arg\max_{\alpha,\beta}\left\{\min_{\boldsymbol{u}}\left\{\sum_{i=1}^{n}\left(u_i + e^{-u_i}\int q_\beta(z|x)\left\{\frac{p_\alpha^{\tau_t}(x,z)}{q_\beta(z|x)}\right\}\mathrm{d}z\right)\right\}\right\} \tag{13}$$

- $q_\beta(z|x)$ 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*

Experiments: Density Estimation

Table: Quantitative evaluation on toy models.

Model	Parameter estimation error † ↓ banana	kidney	rings	river	wave	Likelihood consistency score ↑ banana	kidney	rings	river	wave
MC	3.46	3.9	4.71	1.71	1.78	0.961	0.881	0.508	0.702	0.619
SM	7.79	2.75	3.62	1.64	2.61	×	×	×	×	×
NCE	3.88	2.5	4.81	2.85	**1.20**	0.968	0.882	0.557	0.721	0.759
KEF	×	×	×	×	×	0.973	0.755	0.183	0.436	0.265
DDE	6.59	7.31	24.9	29.1	25.7	0.944	0.830	0.426	0.520	0.186
FML (ours)	**3.05**	**1.9**	**2.59**	**1.13**	1.27	**0.974**	**0.901**	**0.562**	**0.731**	**0.782**

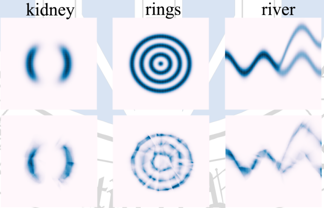


Figure: FML predicted likelihood using nonparametric potentials.

Experiments: Image Models

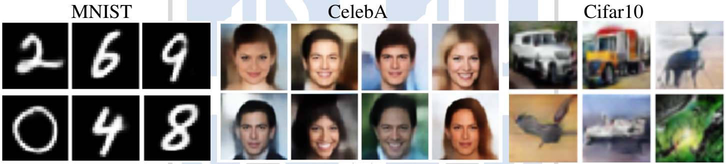


Figure: Synthesized sampled from model trained with FML.

Table: GAN quantitative results.

Cifar10	IS↑	FID↓
GAN	6.29	37.4
DFM	**6.93**	30.7
FML	**6.91**	**30.0**

Table: VAE quantitative results

MNIST	IS↑	FID↓	$-\log \hat{p} \downarrow$
VAE	8.08	24.3	103.7
FML	**8.30**	**22.7**	**101.5**

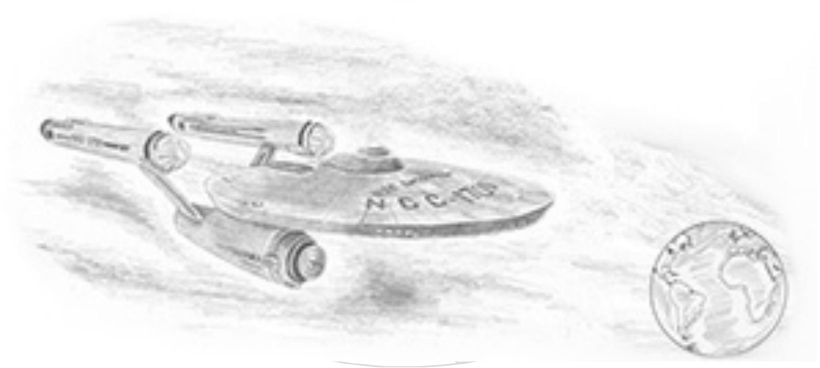
Experiments: Language Models

Table: Results on language models, with the example synthesized text representative of typical results.

	PPL ↓	BLEU-2 ↑	BLEU-3 ↑	BLEU-4 ↑	BLEU-5 ↑
EMNLP WMT news					
VAE	12.5	76.1	46.8	23.1	11.6
FML	**11.6**	**77.2**	**47.4**	**24.3**	**12.2**
MS COCO					
VAE	9.5	82.1	60.7	38.9	24.8
FML	**8.6**	**84.2**	**64.4**	**40.3**	**25.2**
Sampled sentences from respective models on *WMT news*					
VAE	*"China's economic crisis, the number of US exports, which is still in recent years of the UK' s population."*				
FML	*"In addition, police officials have also found a new investigation into the area where they could take a further notice of a similar investigation into."*				

The End

Thank you.



References

- C Tao, et al. On Fenchel Mini-Max Learning. *NeurIPS 2019*

- C Geyer, et al. Markov chain Monte Carlo maximum likelihood. *1991*
- G Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation, 2002*
- A Hyvärinen, et al. Estimation of non-normalized statistical models by score matching. *JMLR 2005*
- M Gutmann, et al. Noise-contrastive estimation: A new estimation principle for unnormalized ... *AISTATS 2010*
- G Alain, et al. What regularized auto-encoders learn from the data-generating distribution. *JMLR 2014*
- Q Liu, et al. Stein variational gradient descent: A general purpose bayesian inference algorithm. *NIPS 2016*
- Y Li, et al. Renyi divergence variational inference. *NIPS 2016.*
- Y Li, et al. Gradient estimators for implicit models. *ICLR 2018*
- D Sutherland, et al. Efficient and principled score estimation with Nystrom kernel exponential families. *2017*
- D Dai, et al. Exponential Family Estimation via Dynamics Embedding. *2017*