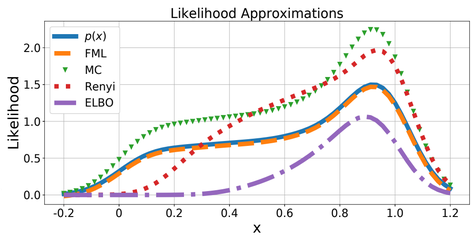
On Fenchel Mini-Max Learning

Chenyang Tao [1] Liqun Chen [1] Shuyang Dai [1] Junya Chen [1,2] Ke Bai [1] Dong Wang [1] Jianfeng Feng [3]
Wenlian Lu [2] Georgiy Bobashev [4] Lawrence Carin [1] Correspondence to: {chenyang.tao, lcarin} @duke.edu
[1] Electrical & Computer Engineering, Duke University, Durham, NC 27708, USA [2] School of Mathematical Sciences, Fudan University, Shanghai 200433, China
[3] Institute of Science and Technology for Brain Inspired Intelligence, Fudan University, Shanghai 200433, China [4] RTI International, Research Triangle Park, NC, USA

NeurIPS | 2019
Vancouver, BC, Canada

Contributions

- A Mini-Max formulation of *Maximal Likelihood Estimation* (MLE), resulting unbiased non-parametric likelihood estimator directly amendable to SGD.
- Amortized likelihood estimation with deep neural networks.
- A latent-variable model competitive to variational inference.
- Our models compare favorably to existing alternatives (see Table A) in likelihood-based distribution learning, both in terms of model estimation and sample generation..



Background

Maximal likelihood estimation. MLE seeks to identify the most probable model parameter $\hat{\theta}_{MLE}$ via maximizing the expected model log-likelihood $\mathcal{L}(\theta) = \frac{1}{n}\sum_i \log p_\theta(x_i)$, where $\{x_i\}$ are empirical observations and p_θ is the normalized density (*i.e.*, $\int p_\theta = 1$). This effectively minimizes the KL-divergence $KL(\hat{p}_d||p_\theta)$ btw the empirical and model distributions.

Challenge of MLE for unnormalized statistical models. For unnormalized model density function $\tilde{p}_\theta(x) = \exp(-\psi_\theta(x))$, where $-\psi_\theta(x)$ is the potential function, the likelihood is given by $p_\theta(x) = \tilde{p}_\theta(x)/Z(\theta)$, here $Z(\theta) = \int e^{-\psi_\theta(x')}dx'$ is known as the partition function and typically does not enjoyed a closed form expression. This makes MLE challenging as any finite-sample estimate for the log-partition will be **biased** due to Jensen's inequality.

$$\log \hat{p}_\theta = -\psi_\theta - \log \hat{Z}_\theta \geq -\psi_\theta - \log Z_\theta = \log p_\theta(x)$$

Variational Inference (VI) for approximate MLE. VI considers maximizing an approximation of the marginal likelihood of a latent variable model $p_\theta(x,z)$, where z is the latent variable to be marginalized. An approximate posterior $q_\phi(z|x)$is introduced to derive a bound to $\log p_\theta(x)$, such as the evidence lower bound $ELBO(x) = \mathbb{E}_{q_\phi(Z|X)}\big[\log p_\theta(x,Z) - \log q_\phi(Z|x)\big] \leq \log p_\theta(x)$

Black-Box (Non-parametric) MLE. To model complex distributions, black-box MLE models are needed in the absence of prior knowledge. Examples include flow-based models (transformations with tractable Jacobian), score matching (2nd order Laplacian), contrastive divergence (MCMC sampling), noise contrastive estimation [1] (contrasting noise proposal) and dynamical dual embedding [2] (tractable dual). All models make some kind of compromise (see Table A).

Table A. Comparison of estimation schemes

Model	Inference	Sampling	Likelihood	Scalability
FML (ours)	Yes	Yes	Estimate	Good
CD	Yes	Yes	No	Good
SM	No	No	No	Poor
NCE	No	No	Estimate	Depends
KEF	No	No	No	Poor
DDE	No	Yes	Exact	Low
VI	Yes	Yes	Bound	Good
Flow	No	Yes	Exact	Tricky
Stein	No	Yes	No	Medium
GAN	No	Yes	No	Good

Fenchel conjugacy. Let $f(t)$ be a be a proper convex, lower-semicontinuous function; then its convex conjugate function $f^*(v)$ is defined as $f^*(v) = \max_{t\in\mathcal{D}(f)}\{tv - f(t)\}$, where $\mathcal{D}(f)$ is the domain of f. f^* is known as the Fenchel conjugate of f, and the pair (f, f^*) is dual to each other. This study exploits the log Fenchel conjugate pair $(-\log t, -1 - \log v)$.

Code available at: https://github.com/chenyang-tao/FML

Algorithm

Empirical data distribution $\hat{p}_d = \{x_i\}_{i=1}^n$
Proposal $q(x)$, learning rate schedule $\{\eta_t\}$
Initialize parameters θ, b
for $t = 1, 2, \cdots$ **do**
  Sample $\{x_{t,j}\}_{j=1}^m \sim \hat{p}_d(x)$, $\{x'_{t,j}\}_{j=1}^m \sim q(x)$
  $u_{t,j} = \psi(x_{t,j}) + b$,
  $I_{t,j} = \exp(\psi_\theta(x_{t,j}) - \psi_\theta(x'_{t,j}) - \log q(x'_{t,j}))$
  $J_t = \sum_j \{u_{t,j} + \exp(-u_{t,j})I_{t,j}\}$
  $[\theta, b] = [\theta, b] - \eta_t \nabla_{[\theta,b]} J_t$
  % Update proposal $q(x)$ if needed
end for

Theory

Mini-max MLE for unnormalized statistical models. For unnormalized statistical model $\log \tilde{p}_\theta(x) = -\psi_\theta(x)$, we have

$$\log p_\theta(x) = -\log\left(\int e^{\psi_\theta(x) - \psi_\theta(x')}dx'\right)$$

Plugging into the Fenchel conjugacy $-\log t = \max_u\{-u - e^{-u}t + 1\}$,

$$-\log p_\theta(x) = \min_{u_x}\{\mathcal{J}(u_x)\}, \mathcal{J}(u_x) \triangleq u_x + e^{-u_x}\int e^{\psi_\theta(x) - \psi_\theta(x')}dx' - 1$$

Note the minimizer $u_x^* = \log p_\theta(x)$. The elimination of log term bypass the biased estimation of partition function with a finite sample estimator. This gives us the *Fenchel Mini-max Learning* (**FML**) scheme of MLE

$$\hat{\theta}_{MLE} = \operatorname{argmax}_\theta\left\{-\min_{\boldsymbol{u}}\{\mathcal{J}(x_i, u_i, \theta)\}\right\}$$

The integral can be more efficiently evaluated with importance sampling.

Gradient analysis of FML. To understand the working of FML, we analyze its gradient

$$\nabla_\theta \mathcal{J}(x, u, \theta) = -\hat{p}_\theta(x)\nabla_\theta\{1/p_\theta(x)\} = \{\hat{p}_\theta(x)/p_\theta(x)\}\nabla_\theta \log p_\theta(x)$$

where $\hat{p}_\theta = e^{u_x}$ is the approximate likelihood and we call $\xi = \hat{p}_\theta(x)/p_\theta(x)$ the approximation error. This observation allows us to establish the convergence result of FML using stochastic gradient descent (SGD).

Proposition. (Convergence of FML-SGD) *Let* $\{\eta_t\}$ *be* the scheduled learning rate and $\{\xi_t\}$ be the approximation error. If the generalized learning rate $\tilde{\eta}_t = \xi_t\eta_t$ satisfies $\sum_t \mathbb{E}[\tilde{\eta}_t] = \infty$ and $\sum_t \mathbb{E}[\tilde{\eta}_t^2] < \infty$, 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* $\hat{\theta}_{MLE}$ *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.* $\|\nabla f(\theta_t)\| \to 0$ *as* $t \to 0$.

FML training of latent variable models. Following a similar derivation, we can arrive at a mini-max training scheme for latent variable models

$$\operatorname{argmax}_\theta\left\{-\min_{\boldsymbol{u}}\left\{\mathbb{E}_{p_d}\left[u_X + \exp(-u_X)\mathbb{E}_{q_\phi(Z|X)}\left[p_\theta(X,Z)/q_\phi(Z|X)\right]\right]\right\}\right\}$$

where $q_\phi(Z|X)$ serves as the proposal (approximate posterior) distribution, with the ground-truth posterior $p_\theta(Z|X)$ as the optimal choice.

Experiments

Likelihood Estimation

Table 1: 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 [36]	7.79	2.75	3.62	1.64	2.61	×	×	×	×	×
NCE [28]	3.88	2.5	4.81	2.85	**1.20**	0.968	0.882	0.557	0.721	0.759
KEF [59]	×	×	×	×	×	0.973	0.755	0.183	0.436	0.265
DDE [16]	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**

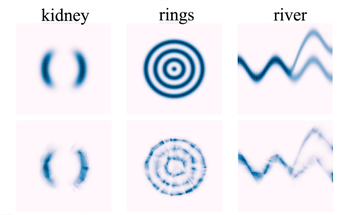


Figure 2: FML predicted likelihood using nonparametric potentials.

Table 2: log-likelihood evaluation on UCI datasets ↑.

Model	wine-red	wine-white	yeast	htru2
KDE	7.74	7.74	3.01	15.47
GMM	7.42	7.97	4.82	22.06
DDE	7.45	7.18	3.79	18.83
FLOW	7.09	7.75	3.31	20.48
NCE	7.29	7.98	4.84	22.05
FML	**8.45**	**8.20**	**4.96**	**22.15**

Generative Modeling

Figure 3: Sampled images from FML-trained models.

Table 3: VAE quantitative results.

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

Table 4: GAN quantitative results.

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

Natural Language Processing

Table 5: 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 popula-tion."*

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."*

Reinforcement Learning

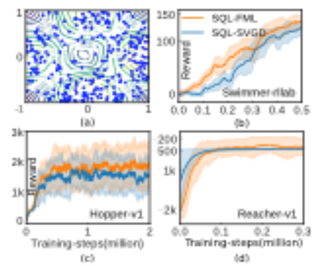


Figure 4: Soft Q-Learning with FML.

References

[1] M Gutmann, et al. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *AISTATS* 2010.

[2] B Dai, et al. Kernel exponential family estimation via doubly dual embedding. *AISTATS* 2019.