ENOT: Expectile Regularization for Fast and Accurate Training of Neural Optimal Transport

Nazar Buzun[1,2], Maksim Bobrin[1,3], Dmitry V. Dylov[1,3]
buzun@airi.net

1. Artificial Intelligence Research Institute
2. Moscow Institute of Physics and Technology
3. The Skolkovo Institute of Science and Technology

$$OT(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \inf_{T:\, T_{\#}\boldsymbol{\alpha}=\boldsymbol{\beta}} \int_{\mathcal{X}} c(x, T(x))\mathrm{d}\boldsymbol{\alpha}(x)$$

In **dual formulation** we optimize sum of Kantorovich potentials f and g

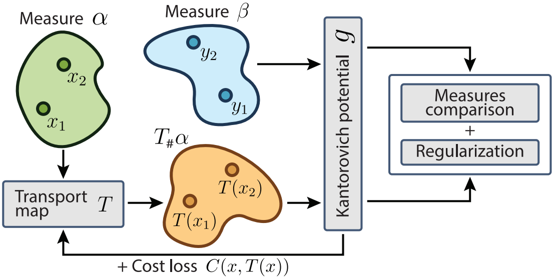
$$\sup_{f\in L_1(\boldsymbol{\alpha}),\, g\in L_1(\boldsymbol{\beta})} \Big[\mathbb{E}_{\boldsymbol{\alpha}}[f(\mathtt{x})] + \mathbb{E}_{\boldsymbol{\beta}}[g(\mathtt{y})]\Big]$$

under condition (c-conjugate dependence)

$$f(\mathtt{x}) = \min_{\mathtt{y}} \{c(\mathtt{x}, \mathtt{y}) - g(\mathtt{y})\}$$

$$g(\mathtt{y}) = \min_{\mathtt{x}} \{c(\mathtt{x}, \mathtt{y}) - f(\mathtt{x})\}$$

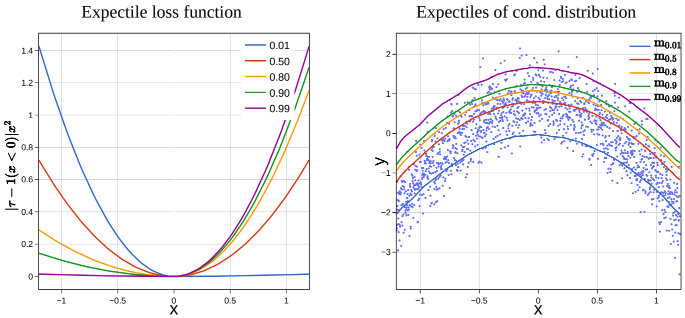
c-conjugate transformation requires computationally intensive fine-tuning



$$\max_{g} \mathbb{E}_{\boldsymbol{\beta}}[g(\mathbf{y})] - \mathbb{E}_{\boldsymbol{\alpha}}[g(T(\mathbf{x}))]$$

$$\min_{T} \mathbb{E}_{\boldsymbol{\alpha}}[c(\mathbf{x}, T(\mathbf{x}))] - \mathbb{E}_{\boldsymbol{\alpha}}[g(T(\mathbf{x}))]$$

Computationally fast but unstable training



$$\min_f \mathbb{E}\mathcal{L}_\tau \left[f(\mathrm{x}) - \mathrm{y}\right], \quad \mathcal{L}_\tau[\mathrm{x}] = \left|\tau - \mathbb{I}(\mathrm{x} < 0)\right| \mathrm{x}^2$$

Particularly when $\tau \to 1$:

$$f(\mathrm{x}) \to \max \text{supp. } \mathrm{y}|\mathrm{x}$$

Desired regularization terms

$$\min_f \mathbb{E} \left[f(\mathrm{x}) - \min_{\mathrm{y}} \{ c(\mathrm{x}, \mathrm{y}) - g(\mathrm{y}) \} \right]^2$$

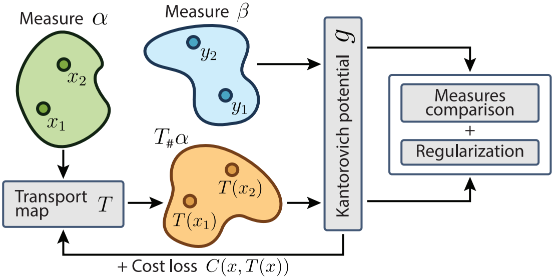
$$\min_g \mathbb{E} \left[g(\mathrm{y}) - \min_{\mathrm{x}} \{ c(\mathrm{x}, \mathrm{y}) - f(\mathrm{x}) \} \right]^2$$

expressed as the expectile regression problem

$$\min_{f,g} \mathbb{E} \mathcal{L}_\tau \left[f(\mathrm{x}) - c(\mathrm{x}, \mathrm{y}) + g(\mathrm{y}) \right]$$

Optimize sum of Kantorovich potentials f and g with expectile regularization

$$\sup_{f \in L_1(\boldsymbol{\alpha}), g \in L_1(\boldsymbol{\beta})} \left[\mathbb{E}_{\boldsymbol{\alpha}}[f(\mathrm{x})] + \mathbb{E}_{\boldsymbol{\beta}}[g(\mathrm{y})] \right] - \lambda \mathbb{E} \mathcal{L}_\tau \left[f(\mathrm{x}) + g(\mathrm{y}) - c(\mathrm{x}, \mathrm{y}) \right]$$



In practical implementation replace $f(\mathrm{x})$ with $\min_T \{c(\mathrm{x}, T(\mathrm{x})) - g(T(\mathrm{x}))\}$

$$\max_g \mathbb{E}_{\boldsymbol{\beta}}[g(\mathrm{y})] - \mathbb{E}_{\boldsymbol{\alpha}}[g(T(\mathrm{x}))] - \lambda \mathbb{E} \mathcal{L}_\tau \left[c(\mathrm{x}, T(\mathrm{x})) - g(T(\mathrm{x})) + g(\mathrm{y}) - c(\mathrm{x}, \mathrm{y})\right]$$

$$\min_T \mathbb{E}_{\boldsymbol{\alpha}}[c(\mathrm{x}, T(\mathrm{x}))] - \mathbb{E}_{\boldsymbol{\alpha}}[g(T(\mathrm{x}))]$$

Method	Conjugate	$D = 64$	$D = 128$	$D = 256$
W2-Cycle	None	7.2	2.0	2.7
MM-Objective	None	8.1	2.2	2.6
MM-R-Objective	None	10.1	3.2	2.7
Monge Gap	None	7.99 $\pm$ 0.19	9.1 $\pm$ 0.29	9.41 $\pm$ 0.21
W2OT-Cycle	None	> 100	> 100	> 100
W2OT-Objective	None	> 100	> 100	> 100
W2OT-Cycle	L-BFGS	> 100	> 100	> 100
W2OT-Objective	L-BFGS	2.08 $\pm$ 0.40	0.67 $\pm$ 0.05	0.59 $\pm$ 0.04
W2OT-Regression	L-BFGS	2.08 $\pm$ 0.39	0.67 $\pm$ 0.05	0.65 $\pm$ 0.07
W2OT-Cycle	Adam	> 100	> 100	> 100
W2OT-Objective	Adam	2.21 $\pm$ 0.32	0.77 $\pm$ 0.05	0.66 $\pm$ 0.07
W2OT-Regression	Adam	2.37 $\pm$ 0.46	0.77 $\pm$ 0.06	0.75 $\pm$ 0.09
ENOT (Ours)	None	0.56 $\pm$ 0.03	0.3 $\pm$ 0.01	0.51 $\pm$ 0.02

$\mathcal{L}_2^{\text{UV}}$ Metric

ENOT Time	None	16 min	21 min	21 min
W2OT-Objective Time	L-BFGS	397 min	571 min	1028 min

Task and image size	CycleGAN	StarGAN	Extr. OT	Ker. OT	ENOT
Handbags ⇒ Shoes 128	23.4	22.36	27.10	26.7	19.19
FFHQ ⇒ Comics 128	-	-	20.95	20.81	17.11
CelebA(f) ⇒ Anime 64	20.8	22.40	14.65	18.28	13.12
CelebA(f) ⇒ Anime 128	-	-	19.44	21.96	18.85

FID Metric

$$\lambda \mathbb{E} \mathcal{L}_\tau \left[c(\mathtt{x}, T(\mathtt{x})) - g(T(\mathtt{x})) + g(\mathtt{y}) - c(\mathtt{x}, \mathtt{y}) \right]$$

- **Theoretical Guarantees:**
  - Unbiased estimation when $\tau \to 1$,
  - Forces potential g to be c-concave
- **Empirical Advantages:**
  - 3-10x faster convergence,
  - Improved stability,
  - No extensive hyperparameters tuning (λ and τ)
- **Applications:**
  - Various cost functions beyond W2 distance

AIRI

МФТИ

Skoltech

@airi_research_institute

skylooop.github.io/enot

ott-jax.readthedocs.io