IMPROVED STOCHASTIC OPTIMIZATION OF LOGSUMEXP

Anonymous authors
Paper under double-blind review

ABSTRACT

The LogSumExp function, also known as the free energy, plays a central role in many important optimization problems, including entropy-regularized optimal transport and distributionally robust optimization (DRO). It is also the dual to the Kullback-Leibler (KL) divergence, which is widely used in machine learning. In practice, when the number of exponential terms inside the logarithm is large or infinite, optimization becomes challenging since computing the gradient requires differentiating every term. Previous approaches that replace the full sum with a small batch introduce significant bias. We propose a novel approximation to LogSumExp that can be efficiently optimized using stochastic gradient methods. This approximation is rooted in a sound modification of the KL divergence in the dual, resulting in a new f-divergence called the *safe KL divergence*. The accuracy of the approximation is controlled by a tunable parameter and can be made arbitrarily small. Like the LogSumExp, our approximation preserves convexity. Moreover, when applied to an L-smooth function bounded from below, the smoothness constant of the resulting objective scales linearly with L. Experiments in DRO and continuous optimal transport demonstrate the advantages of our approach over state-of-the-art baselines and the effective treatment of numerical issues associated with the standard LogSumExp and KL.

1 INTRODUCTION

Optimization problems arising in various fields involve the LogSumExp function, or, more generally, the log-partition functional

$$F(\varphi;\mu) \coloneqq \ln \int e^{\varphi(x)}\,\mathrm{d}\mu(x) \in (-\infty,\infty] \tag{1}$$

mapping a measurable function φ to $(-\infty,\infty]$ based on a probability measure μ. The goal in such optimization problems is to minimize an objective involving F w.r.t. φ over some class.

LogSumExp function appears commonly in optimization objectives, e.g., multiclass classification with softmax probabilities (Bishop & Nasrabadi, 2006), semi-dual formulation of entropy-regularized optimal transport (OT) (Peyré & Cuturi, 2019; Genevay et al., 2016), minimax problems (Pee & Royset, 2011), distributionally robust optimization (DRO) (Hu & Hong; Ben-Tal et al., 2013; Kuhn et al., 2024), maximum likelihood estimation (MLE) for exponential families and graphical models (Wainwright et al., 2008), variational Bayesian methods (Khan & Nielsen, 2018; Khan & Rue, 2023), information geometry (Amari & Nagaoka, 2000), KL-regularized Markov decision processes (Tiapkin et al., 2024). These problems involve minimizing $F(\varphi;\mu)$ w.r.t. a function φ, potentially parameterized by a vector θ, e.g., a vector of neural network weights. Such optimization is characterized by two challenges. First, the decision variable φ or θ often has large or infinite dimension. Second, the support of the measure μ can also be large or infinite. The first challenge is usually addressed by the use of first-order methods like Stochastic Gradient Descent (SGD), which are suitable for high-dimensional problems due to their cheap iterations. Unfortunately, there is no universal way to tackle the second challenge. If the number of exponential terms under the logarithm is large or infinite, i.e., the Monte Carlo estimate $\log \sum_{i=1}^{N} e^{\varphi(x_i)}$ for a large N, the gradient computation requires differentiating each term, and, to the best of our knowledge, no cheap unbiased stochastic gradient have been proposed.

In the current work, we propose a general-purpose approach to tackle both mentioned challenges. To that end, we use a SoftPlus approximation of $F(\varphi; \mu)$ that allows using stochastic gradient methods while remaining close to the original objective. We start with a variational formulation analogous to the one in the Gibbs principle, but with the KL-divergence replaced with another f-divergence – the *safe KL* divergence. The corresponding variational problem can be of interest itself, as it possesses some properties which can be beneficial compared to the KL penalty – e.g., uniform density bound. Moreover, it can be also viewed as an approximation of a conditional value at risk functional (CVaR). In fact, the same functional (with different parameters) appeared in Soma & Yoshida (2020) in the context of smooth CVaR approximation. Thus, we demonstrate that it generates a family of problems including CVaR and LogSumExp minimization as limit cases.

Related works. We are not aware of any universal way to efficiently treat optimization objectives involving the LogSumExp functional (1), especially in infinite-dimensional settings. This functional appeared in different applications and was treated on a case-by-case basis. Bouchard (2007) studies three upper bounds on LogSumExp for approximate Bayesian inference. One of them is a particular case of the approximation proposed in the present work. Titsias (2016) constructs a bound on softmax probabilities and shows that it leads to a bound on LogSumExp in the context of multiclass classification. Nielsen & Sun (2016) approximate LogSumExp in the context of estimating divergences between mixture models. Their approximation combines LogSumExp bounds based on $\min$ and $\max$. Tucker et al. (2017); Luo et al. (2020) propose and study unbiased estimators for latent variable models based on Russian Roulette truncation. Lyne et al. (2015); Spring & Shrivastava (2017) focus on estimating the partition function itself, and do not consider questions of optimization involving the partition function. Hu & Hong and, susequently, Levy et al. (2020) study DRO problems with f-divergences. They propose a batch-based approximation. When the ambiguity set is the unit simplex, and KL divergence penalty is used, the original objective is the LogSumExp of the losses over the entire dataset, while the approximation replaces it with the average of LogSumExp terms computed on individual batches. This approximation introduces bias, which can only be reduced by using large batch sizes. Deterministic LogSumExp *maximization* and minimization were considered in Selvi et al. (2020) and Kan et al. (2023), respectively. For stabilizing numerics related to evaluation of LogSumExp function, we refer to Blanchard et al. (2021); Higham (2021).

Contributions. Our main contributions are as follows:

1. We introduce a general-purpose and computationally efficient approach for handling the LogSumExp function in large-scale optimization problems by proposing a novel relaxation of this function. The proposed relaxation preserves key properties of the original LogSumExp function, such as convexity and smoothness, and enables the use of stochastic gradient methods for machine learning tasks. Furthermore, our method only requires a simple and tunable scalar parameter, allowing the relaxation to be made arbitrarily close to the original LogSumExp objective as desired.
2. We provide the theoretical backbone of this approximation, demonstrating that it is due to a modified version of the KL-divergence in the dual formulation. We termed the resulting f-divergence the *(Overflow-)Safe KL* divergence. It can be applied to various applications where KL-divergence is used.
3. We empirically demonstrate the effectiveness of our approach on tasks, including computing continuous entropy-regularized OT and various DRO formulations. Our method outperforms existing state-of-the-art baselines in these applications and circumvents the overflow issue (Remark 3.1). It can also be combined with existing techniques. Therefore, it serves as a versatile tool for solving large-scale optimization problems.
4. Additionally, we provide insights into a few remarkable connections between the proposed approximation and existing notions such as the conditional value-at-risk.

Notation. Given $a, a_1, \dots, a_n \in \mathbb{R}$, we define $\mathrm{LogSumExp}(a_1, \dots, a_n) \coloneqq \log(\sum_{i=1}^n e^{a_i})$ and $\mathrm{SoftPlus}(a) \coloneqq \log(1 + e^a)$. Given a measurable space $\mathcal{X}$, by $\mathcal{P}(\mathcal{X})$ we denote the space of probability measures on $\mathcal{X}$, and by $\mathcal{C}(\mathcal{X})$ the space of continuous functions on $\mathcal{X}$. Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$. Define Kullback–Leibler (KL) divergence as

$$D_{KL}(\mu, \nu) \coloneqq \begin{cases} \int_{\mathcal{X}} \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x)\, \mathrm{d}\mu(x) & \mu \ll \nu, \\ +\infty & \text{otherwise}, \end{cases}$$

where $\log$ is the natural logarithm and $\mu \ll \nu$ denotes that μ is absolutely continuous w.r.t. ν.

2 SOFTPLUS APPROXIMATION OF LOG-PARTITION FUNCTION

In this section, we present our approximation to the log-partition function (1) and describe its theoretical properties. Recall that by the Gibbs variational principle

$$F(\varphi;\mu) = \sup_{\nu}\left\{\int_{\mathcal{X}} \varphi(x)\,\mathrm{d}\nu(x) - D_{KL}(\nu,\mu) : \nu \in \mathcal{P}(\mathcal{X}),\ \int_{\mathcal{X}} |\varphi(x)|\,\mathrm{d}\nu(x) < \infty\right\} \tag{2}$$

with the maximum attained at the Gibbs measure $d\nu^*(x) = e^{\varphi(x)-F(\varphi;\mu)}d\mu(x)$, once $F(\varphi;\mu) < \infty$, see (Gibbs, 1902, Chapter XI, Theorem VI) or (Polyanskiy & Wu, 2025, Proposition 4.7) for the modern treatment.

We are going to construct an approximation of F with better regularity properties by changing D_{KL} to another f-divergence. Specifically, for any $0 < \rho < 1$, let us define the following.

Definition 2.1 (Safe KL entropy). *We define the safe KL entropy generator $f_\rho\colon [0,\infty) \to \mathbb{R}$ by*

$$f_\rho(t) \coloneqq \begin{cases} t\log t + 1 + \frac{1-\rho t}{\rho}\log(1-\rho t), & 0 \le t \le \frac{1}{\rho}, \\ +\infty, & \textit{otherwise}. \end{cases} \tag{3}$$

The resulting f_ρ-divergence, which we refer to as the safe KL divergence*, is given by*

$$D_\rho(\nu,\mu) \coloneqq \begin{cases} \int_{\mathcal{X}} f_\rho\left(\frac{d\nu}{d\mu}(x)\right)\,\mathrm{d}\mu(x), & \nu \ll \mu, \\ +\infty, & \textit{otherwise}. \end{cases} \tag{4}$$

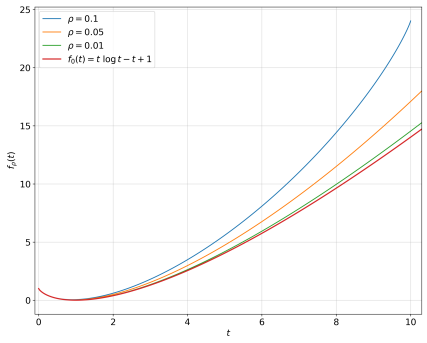

Figure 1: $f_\rho(t)$ for different values of ρ.

It is easy to see that $f_\rho(t) \to f_0(t) \coloneqq t\log t + 1 - t$ as $\rho \to 0$. Since f_0 induces the standard KL-divergence, D_ρ is its approximation with accuracy regulated by the parameter ρ.

Using the variational representation, we define

$$F_\rho(\varphi;\mu) \coloneqq \sup_{\nu}\left\{\int_{\mathcal{X}} \varphi(x)\,\mathrm{d}\nu(x) - D_\rho(\nu,\mu) : \nu \in \mathcal{P}(\mathcal{X}),\ \int_{\mathcal{X}} |\varphi(x)|\,\mathrm{d}\nu(x) < \infty\right\}. \tag{5}$$

(i.e., $F_\rho(\cdot;\mu)$ is the convex conjugate of $D_\rho(\cdot,\mu)$). Note that the last term in f_ρ prevents the density $\frac{d\nu}{d\mu}$ from being too large. In particular, it can not be greater than $\frac{1}{\rho}$. This can make the safe KL divergence a reasonable choice for unbalanced OT or DRO, as it imposes a hard constraint on the reweighting unlike the standard D_{KL}. Moreover, it can also be used instead of the entropy penalization in regularized OT (cf. capacity constrained transport in (Benamou et al., 2015, section 5.2)).

Again, by the convex duality and the variational principle (see Birrell et al., 2022, Theorem 6), we state the following properties.

Lemma 2.2. *The functional F_ρ defined by* (5) *has an equivalent variational representation*

$$F_\rho(\varphi;\mu) = \inf_{\alpha\in\mathbb{R}} \alpha + \int_{\mathcal{X}} f_\rho^*(\varphi(x) - \alpha)\,\mathrm{d}\mu(x).$$

It is straightforward to check the following.

Lemma 2.3. *The conjugate function to f_ρ is a rescaled SoftPlus, specifically,*

$$f_\rho^*(s) \coloneqq \sup_{t\in\mathbb{R}_+} st - f_\rho(t) = \frac{1}{\rho}\log\left(1 + \rho e^s\right) - 1.$$

Therefore, we obtain

$$F_\rho(\varphi;\mu) = \inf_{\alpha\in\mathbb{R}} \alpha - 1 + \frac{1}{\rho}\int_{\mathcal{X}} \log\left(1 + \rho e^{\varphi(x)-\alpha}\right)\,\mathrm{d}\mu(x). \tag{6}$$

In essence, we have replaced the exponential function with a rescaled SoftPlus. Furthermore, it is easy to see that the optimal α^* satisfies

$$\int_{\mathcal{X}} \frac{e^{\varphi(x)-\alpha^*}}{1+\rho e^{\varphi(x)-\alpha^*}}\,\mathrm{d}\mu(x) = 1, \tag{7}$$

in particular, $\alpha^* < F(\varphi;\mu)$. Moreover, the maximum in (5) is attained at $d\nu_\rho^*(s) = \frac{e^{\varphi(x)-\alpha^*}}{1+\rho e^{\varphi(x)-\alpha^*}}\,\mathrm{d}\mu(x)$. Note that $0 < \frac{d\nu_\rho^*(x)}{d\mu(x)} < \frac{1}{\rho}$, which is due to the fact that the derivative of $t\log t$ explodes at 0, preventing reaching the constraint.

The next proposition (proved in Appendix A) ensures that F_ρ is a valid approximation of F.

Proposition 2.4. *Let $\mu \in \mathcal{P}(\mathcal{X})$ and φ be a measurable function on $\mathcal{X}$.*

(i) For all $0 < \rho \le \rho' < 1$, it holds $F_{\rho'}(\varphi;\mu) \le F_\rho(\varphi;\mu)$.
(ii) As $\rho \to 0+$, $F_\rho(\varphi;\mu) \to F_0(\varphi;\mu) := F(\varphi;\mu)$.
(iii) If $F(2\varphi;\mu) < \infty$, then for all $0 < \rho \le \frac{1}{4}e^{2F(\varphi;\mu)-F(2\varphi;\mu)}$

$$F_\rho(\varphi;\mu) \ge F(\varphi;\mu) + \frac{\rho}{2} - 4\rho e^{F(2\varphi;\mu)-2F(\varphi;\mu)}. \tag{8}$$

(iv) If $\varphi(x) \le M$ for all $x \in \mathcal{X}$, then $F_\rho(\varphi;\mu) \ge F(\varphi;\mu) - \rho e^{M-F(\varphi;\mu)}$ for $\rho \in \left(0, e^{F(\varphi;\mu)-M}\right)$.

In particular, (i) and (iii) show that $F_\rho - O(\rho) \le F \le F_\rho$, and thus the parameter ρ allows one to control the approximation accuracy. In the case of LogSumExp, the above proposition yields the following simple bounds.

Corollary 2.5. *Let $a_1, \dots, a_n \in \mathbb{R}$. Then for any $0 < \rho < 1$*

$$\mathrm{LogSumExp}(a_1,\dots,a_n) - \rho \le \inf_{\alpha\in\mathbb{R}} \alpha - 1 + \frac{1}{\rho}\sum_{i=1}^n \log(1+\rho e^{a_i-\alpha}) \le \mathrm{LogSumExp}(a_1,\dots,a_n).$$

For $\rho = 1$ our approximation coincides with Bouchard's bound for LogSumExp (Bouchard, 2007).

2.1 Links to CVaR

Recall that the conditional value at risk (CVaR) w.r.t. a probability measure $\mu \in \mathcal{P}(\mathcal{X})$ at level $\rho \in (0,1)$, associated with a function φ, can be defined (in the case of continuous distribution) as

$$\mathrm{CVaR}_\rho(\varphi;\mu) := \mathbb{E}_{X\sim\mu}\left[\varphi(X) | \varphi(X) \ge Q_{1-\rho}\right] = \frac{1}{\rho}\int_{\varphi(x)\ge Q_{1-\rho}} \varphi(x)\,\mathrm{d}\mu(x),$$

where $Q_{1-\rho}$ is the $(1-\rho)$-quantile of $\varphi(X)$, $X \sim \mu$ (Rockafellar et al., 2000). Moreover, by Theorem 1 in Rockafellar et al. (2000) CVaR also has the following variational formulation:

$$\mathrm{CVaR}_\rho(\varphi;\mu) = \inf_{\alpha\in\mathbb{R}} \alpha + \frac{1}{\rho}\int_{\mathcal{X}} (\varphi(x)-\alpha)_+\,\mathrm{d}\mu(x). \tag{9}$$

Remarkably, in Soma & Yoshida (2020) the authors obtained a smooth approximation to CVaR which, up to an additive constant, has the same form as F_ρ. However, they considered the approximation w.r.t. a different parameter—a "temperature" inside SoftPlus. Finally, Levy et al. (2020) proposed another similar smoothed version of CVaR (KL-regularized CVaR) in the context of DRO. For our approximation, we obtain the following bounds.

Proposition 2.6. *For all $0 < \rho < 1$ and $\lambda > 0$*

$$\mathrm{CVaR}_\rho(\varphi;\mu) + \lambda(\log\rho - 1) \le \lambda F_\rho(\varphi/\lambda;\mu) \le \mathrm{CVaR}_\rho(\varphi;\mu) + \lambda\left(\log\rho - 1 + \frac{1}{\rho}\right). \tag{10}$$

2.2 The case of parametric models

In some applications, the function φ is defined as the parametric loss function $L(x,\theta)$ and the goal is to minimize objective involving (1) w.r.t. parameter θ to find the best model from the parametric

family. In this section, we study our approximation to (1) in this parametric setting. Fix some closed parameter set $\Theta \subset \mathbb{R}^d$ and a loss function $L\colon \mathcal{X} \times \Theta \to \mathbb{R}$. Combining our approximation (6) and the minimization w.r.t. parameter θ, we obtain the following minimization problem (note that we shifted α by $\log \rho$ compared to (6))

$$\min_{\theta\in\Theta,\alpha\in\mathbb{R}} G_\rho(\theta,\alpha) \coloneqq \alpha + \log\rho - 1 + \frac{1}{\rho}\int_{\mathcal{X}} \log\left(1 + e^{L(x,\theta)-\alpha}\right) \mathrm{d}\mu(x).$$

Clearly, G_ρ is convex in α. Moreover, if L is convex in θ for μ-a.e. x, then G_ρ is jointly convex, meaning that our approximation preserves convexity.

Note that $f_\rho(t) = \frac{1}{\rho}\left((\rho t)\log(\rho t) + (1-\rho t)\log(1-\rho t)\right) + 1 - t\log\rho$. Thus, unlike the KL entropy function $t\log t + 1 - t$, f_ρ possesses the following favorable properties:

Lemma 2.7. *The entropy function f_ρ is ρ-strongly convex. Its conjugate function f_ρ^* is $\frac{1}{\rho}$-smooth.*

The above properties are important from the computational optimization point of view. Recall that $\frac{d}{dt}\log(1+e^t) = \frac{e^t}{1+e^t} =: \sigma(t)$. Thus, we immediately obtain the following formulas for the gradient:

$$\nabla_\theta G_\rho(\theta,\alpha) = \frac{1}{\rho}\int_{\mathcal{X}} \sigma\left(L(x,\theta)-\alpha\right)\nabla_\theta L(x,\theta)\,\mathrm{d}\mu(x),$$
$$\partial_\alpha G_\rho(\theta,\alpha) = 1 - \frac{1}{\rho}\int_{\mathcal{X}} \sigma\left(L(x,\theta)-\alpha\right)\,\mathrm{d}\mu(x).$$

This yields, in particular, that the variance of the (naïve) stochastic gradient is bounded by $\frac{1}{\rho}$ and the second moment of $\nabla_\theta L(X,\theta)$, $X \sim \mu$. In the same way one can calculate the Hessian of G_ρ, see Appendix C. By Proposition C.2, if $L(x,\theta)$ is bounded from below, then G_ρ is smooth on $\Theta \times (-\infty, a]$ for any $a \in \mathbb{R}$ meaning that our approximation preserves smoothness of the loss L.

3 Applications

In this section we consider several particular applications involving the objective (1) and show numerically, that our general-purpose approach based on approximation (6) leads to better performance of SGD-type algorithms than the baseline algorithms designed specifically for these applications. Source code for all experiments can be found in supplementary material.

3.1 Continuous entropy-regularized OT

The classical optimal transport (Monge–Kantorovich) problem consists in finding a coupling of two probablility measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$ which minimizes the integral of a given measurable cost function $c\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}_+$ (e.g., a distance), i.e., $W(\mu,\nu) \coloneqq \inf_{\pi\in\Pi(\nu,\mu)} \int c(x,z)\,\mathrm{d}\pi(x,z)$, where $\Pi(\mu,\nu) \subset \mathcal{P}(\mathcal{X}\times\mathcal{X})$ is the set of couplings (transport plans) of μ and ν (see Kantorovich, 1942; Villani, 2008; Santambrogio, 2015). For simplicity of demonstration, we assume that the measures are defined on the same space $\mathcal{X}$, but the results extend trivially to the case of two different spaces. Following Cuturi (2013), we consider entropy-regularized optimal transport (eOT) problem:

$$\min_{\pi\in\Pi(\nu,\mu)} \int_{\mathcal{X}\times\mathcal{X}} c(x,z)\,\mathrm{d}\pi(x,z) + \varepsilon D_{KL}(\pi, \nu\otimes\mu) \tag{11}$$

where $\nu\otimes\mu$ is the product measure. It is known that eOT admits the following dual and semi-dual formulations (see, e.g., Genevay et al. (2016)):

$$W_\varepsilon(\mu,\nu) = \underbrace{\max_{u,v\in\mathcal{C}(\mathcal{X})} \iint_{\mathcal{X}\times\mathcal{X}} f_\varepsilon(x,y,u,v)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)}_{\text{dual}} = \underbrace{\max_{v\in\mathcal{C}(\mathcal{X})} \int_{\mathcal{X}} h_\varepsilon(x,v)\,\mathrm{d}\mu(x)}_{\text{semi-dual}},$$

where

$$f_\varepsilon(x,y,u,v) \coloneqq u(x) + v(y) - \varepsilon\exp\left(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\right), \tag{12}$$

$$h_\varepsilon(x,v) \coloneqq \int_{\mathcal{X}} v(y)\mathrm{d}\nu(y) - \varepsilon\log\left(\int_{\mathcal{X}} \exp\left(\frac{v(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}\nu(y)\right) - \varepsilon, \tag{13}$$

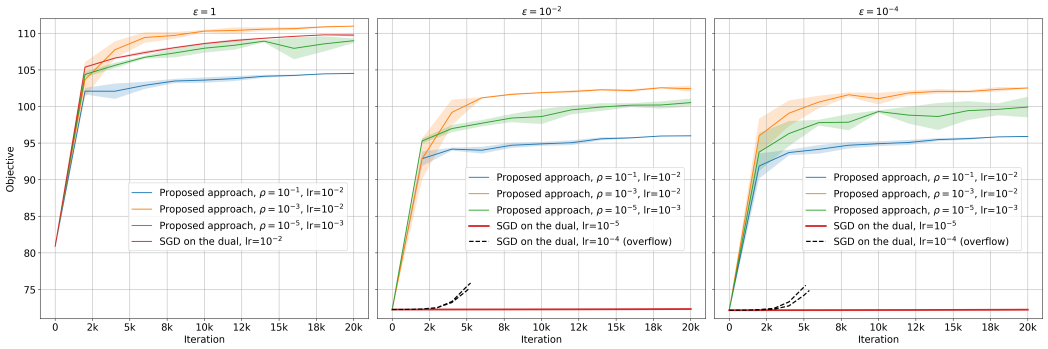

Figure 2: Test-set eOT semi-dual objective vs. iteration for different regularization strengths ε (left to right: 1, 10^{-2}, 10^{-4}). Lines show the mean across 5 runs; shaded areas are $\pm$ one standard deviation. We compare LSOT (red) with our method (colored by ρ). Dashed black curves are examples where LSOT with lr=10^{-4} terminates early due to overflow, while lr=10^{-5} results in a prohibitively slow convergence (nearly horizontal red lines for $\varepsilon = 10^{-2}, 10^{-4}$). Our proposed method remains stable and efficient for all ε.

and $\varepsilon > 0$ is the regularization coefficient. In the LSOT framework (Seguy et al., 2018), the potentials u and v are parameterized by neural networks and optimized via SGD. While Appendix B.1 contains a more detailed literature review, we briefly position LSOT among other solvers to motivate its selection as a baseline. LSOT offers two key advantages relevant to our goals: it is **less computationally intensive** than modern solvers requiring adversarial training (Korotin et al., 2023; Gushchin et al., 2023; Asadulaev et al., 2024) or iterative Langevin dynamics (Mokrov et al., 2024), and it supports a **general cost function**—contrary to other efficient solvers like (Korotin et al., 2024) tailored to the quadratic cost. Therefore, to solve eOT with a general cost function under modest computational constraints, we adopt the LSOT framework as our primary baseline. In Appendix B.2 we compare also to (Genevay et al., 2016) who use an RKHS parametrization for the potentials u and v.

Remark 3.1 (The overflow issue)**.** *The main drawback of this approach is the presence of the exponent in the dual objective (and consequently in the SGD updates). Specifically, exponents are prone to floating-point exceptions (Goldberg, 1991), especially if the regularization parameter ε is relatively small, which is often the case. For example, if $\varepsilon = 0.01$ and $z \geq 7.1$, then $e^{z/\varepsilon}$ exceeds the representable range of a double-precision (float64) floating-point number — an overflow happens. When single precision (float32) is used, an overflow happens even for $z \geq 0.89$.*

Our approach. If we consider instead the semi-dual formulation and use the approximation (6), we arrive at the problem

$$\max_{v,\alpha \in \mathcal{C}(\mathcal{X})} \iint_{\mathcal{X}\times\mathcal{X}} \tilde{h}_\varepsilon(x, y, v, \alpha)\, \mathrm{d}\mu(x)\, \mathrm{d}\nu(y) \tag{14}$$

$$\text{with } \tilde{h}_\varepsilon(x, y, v, \alpha) \coloneqq v(y) - \alpha(x) - \frac{\varepsilon}{\rho} \log\big(1 + \rho e^{(v(y) - c(x,y) - \alpha(x))/\varepsilon}\big) - \varepsilon, \tag{15}$$

which also admits neural network parameterization and optimization via SGD. One can show, in the same way as in Genevay et al. (2016), that this corresponds to the regularized OT problem (11) with *Safe KL divergence* D_ρ rather than the usual KL, i.e.

$$\min_{\pi \in \Pi(\nu,\mu)} \int_{\mathcal{X}\times\mathcal{X}} c(x, z)\, \mathrm{d}\pi(x, z) + \varepsilon D_\rho(\pi, \nu \otimes \mu).$$

Note that this problem, in turn, can be viewed as a combination of the entropy-regularized and the capacity-constrained optimal transport. For $\rho > 0$, this approach is much more stable than the previous one when used in SGD. We illustrate this in the following experiments.

Experiments. We consider the MNIST (Deng, 2012) and EMNIST-letters (Cohen et al., 2017) datasets as samples from the distributions μ (digits) and ν (letters). Manhattan distance ℓ_1 is chosen as the cost function for computing eOT between μ and ν. We parameterize the functions u, v in

LSOT and v, α in our proposed approach using a multilayer perceptron with two hidden layers (dimensions 256 and 128) and ReLU activations. The batch size is 256, and the learning rate is selected via grid search over $\{10^{-6}, 10^{-5}, \dots, 10^{-1}\}$. The objective is evaluated on the empirical distributions of the dedicated test sets.

Figure 2 shows the performance of LSOT with the best learning rate for each regularization parameter $\varepsilon \in \{1, 10^{-2}, \dots, 10^{-4}\}$. It also depicts our proposed approach with the best learning rate for each $\rho \in \{10^{-1}, 10^{-3}, 10^{-5}\}$. The baseline performs adequately under strong regularization ($\varepsilon = 1$). However, for weaker regularization, a learning rate of 10^{-5} is required to avoid numerical instability, which leads to prohibitively slow progress (red curves). Increasing the rate to 10^{-4} (dashed black curves) results in numerical overflow after only ≈5k iterations, forcing us to abort the LSOT runs at that point.

Performance of our proposed approach align with the theoretical analysis in Section 2. A large ρ yields stable convergence but introduces an approximation gap, while a very small ρ degrades smoothness, necessitating a smaller step size and slower training. The intermediate value $\rho = 10^{-3}$ achieves the best trade-off, providing both accuracy and sufficient smoothness. In summary, our proposed approach to eOT is computationally efficient, accommodates general costs, and handles weak regularization robustly, thereby overcoming a key limitation of LSOT.

3.2 Distributionally robust optimization with KL divergence

One of the approaches to training a model that is robust to data distribution shifts and noisy observations is called Distributionally Robust Optimization (DRO) (Kuhn et al., 2024). In contrast to the standard Empirical Risk Minimization (ERM) approach, which minimizes the average loss on the training sample, DRO minimizes the risk for the worst-case distribution among those close to a reference measure (e.g., empirical distribution). A prominent example is KL divergence DRO (Hu & Hong), which is formulated as the saddle-point problem

$$\min_{\theta\in\Theta} \max_{p\in\Delta^n} \sum_{i=1}^n p_i \ell_i(\theta) - \lambda D_{KL}(p, \hat{p}), \tag{16}$$

where $\theta \in \Theta$ is the model parameters, $\ell_i(\theta)$ is the respective loss on the i-th training example, Δ^n is the unit simplex in $\mathbb{R}^n$, $\hat{p} \in \Delta^n$ is the weight vector defining the empirical distribution (typically $\hat{p} = \frac{1}{n}\mathbf{1}$), and D_{KL} is the Kullback–Leibler divergence which discourages distributions that are too far from the empirical one, $\lambda > 0$ is the penalty coefficient. For fixed θ, the solution of the maximization problem is given by $p_i^*(\theta) := \frac{e^{\ell_i(\theta)/\lambda}}{\sum_j e^{\ell_j(\theta)/\lambda}}$, which reduces the problem to

$$\min_{\theta\in\Theta} \mathcal{L}(\theta) := \lambda \log\Big(\frac{1}{n}\sum_{i=1}^n e^{\ell_i(\theta)/\lambda}\Big). \tag{17}$$

However, when n is large, computing the full gradient $\nabla\mathcal{L}(\theta) = \sum_{i=1}^n p_i^*(\theta)\nabla\ell_i(\theta)$ becomes costly. A straightforward approach (Levy et al., 2020) is to sample a batch D, compute the respective softmax weights $p_i^D(\theta) := \frac{e^{\ell_i(\theta)/\lambda}}{\sum_{j\in D} e^{\ell_j(\theta)/\lambda}}$, and define a gradient estimator by

$$\tilde{\nabla}_D\mathcal{L}(\theta) = \sum_{i\in D} p_i^D(\theta)\nabla\ell_i(\theta). \tag{18}$$

However, this introduces a bias and requires using large batch sizes to keep it sufficiently small.

Our approach. Instead, we propose to use the approximation (6), which results in the problem

$$\min_{\substack{\theta\in\Theta\\ \alpha\in\mathbb{R}}} G(\theta,\alpha) := \frac{1}{n}\sum_{i=1}^n \Big\{\alpha + \frac{\lambda}{\rho}\log\big(1 + \rho e^{(\ell_i(\theta)-\alpha)/\lambda}\big)\Big\}. \tag{19}$$

Like in the previous subsection, this can be interpreted as switching from D_{KL} penalty in (16) to *Safe KL* D_ρ. The respective gradient estimators are

$$\tilde{\nabla}_\theta^D G(\theta,\alpha) := \frac{1}{|D|}\sum_{i\in D}\sigma_\rho\left(\tfrac{\ell_i(\theta)-\alpha}{\lambda}\right)\nabla\ell_i(\theta), \quad \tilde{\nabla}_\alpha^D G(\theta,\alpha) := 1 - \frac{1}{|D|}\sum_{i\in D}\sigma_\rho\left(\tfrac{\ell_i(\theta)-\alpha}{\lambda}\right). \tag{20}$$

Table 1: Objective value (17) (mean ± std across 10 runs) at epoch 50 for baseline (18) (Levy et al., 2020) and proposed gradient estimator (20) with different ρ values. Results are shown for various penalty coefficients λ and batch sizes $|D|$, with optimal learning rates selected from $\{10^{-9}, \dots, 10^{-4}\}$. Best results per column are shown in bold.

Approach	$\lambda = 1/5$			$\lambda = 1$			$\lambda = 5$		
	$\lvert D\rvert = 10$	$\lvert D\rvert = 10^2$	$\lvert D\rvert = 10^3$	$\lvert D\rvert = 10$	$\lvert D\rvert = 10^2$	$\lvert D\rvert = 10^3$	$\lvert D\rvert = 10$	$\lvert D\rvert = 10^2$	$\lvert D\rvert = 10^3$
Baseline (18)	26.9±0.7	15.6±6.0	**9.1±4.6**	20.0±0.9	5.2±2.9	**2.3±0.2**	0.87±0.01	0.88±0.00	0.79±0.00
(20), $\rho = 10^{-1}$	27.7±0.6	27.7±0.7	40.1±0.5	21.1±1.1	21.3±1.1	21.8±2.1	0.87±0.01	0.87±0.01	0.88±0.02
(20), $\rho = 10^{-3}$	21.2±9.8	18.6±7.7	25.3±0.1	**2.1±0.0**	**2.1±0.0**	**2.5±1.2**	**0.76±0.02**	**0.78±0.00**	**0.78±0.00**
(20), $\rho = 10^{-5}$	19.2±9.6	17.5±6.6	24.3±0.3	3.0±0.0	3.0±0.0	3.0±0.0	1.03±0.00	1.03±0.00	1.03±0.00

Experiments. Consider the California housing dataset (Pace & Barry, 1997) consisting of 20,640 objects represented by 8 features. Let ℓ_i be the squared error of a linear model, $\ell_i(\theta) = (y_i - \theta^\top x_i)^2$. We use accelerated SGD with the gradient estimator (18) (Levy et al., 2020) as the baseline approach for solving (17), and compare it to our proposed gradient estimator (20). We consider various values of the penalty coefficient $\lambda \in \{1/5, 1, 5\}$ and batch sizes $|D| \in \{10, 10^2, 10^3\}$. For each configuration, we select the optimal learning rate from $\{10^{-9}, 10^{-8}, \dots, 10^{-4}\}$. The approximation accuracy parameter ρ in our method is varied across $\{10^{-1}, 10^{-3}, 10^{-5}\}$. Momentum is fixed at 0.9 (without tuning), and the least squares solution is used as the initial point for optimization.

Numerical results are presented in Table 1, showing the objective value (mean ± standard deviation across 10 runs) after 50 epochs, where the methods typically reach a plateau. In each column, the best-performing configurations are highlighted in bold. For $\lambda = 1/5, |D| \in \{10, 10^2\}$, no results are displayed in bold as all configurations perform similarly. As seen from the table, the baseline and our estimator achieve comparable performance for large batch sizes ($|D| = 10^3$). However, for smaller batches, our method typically outperforms the baseline. Both approaches handle various λ values well, with the exception of the baseline method combined with small batch sizes.

Regarding the approximation parameter ρ, large values ($\rho = 10^{-1}$) generally result in a noticeable approximation gap, while excessively small values ($\rho = 10^{-5}$) deteriorate the smoothness of the objective and consequently slow convergence. The intermediate value $\rho = 10^{-3}$ thus provides the best trade-off in this experiment, offering both good approximation accuracy and favorable optimization properties.

3.3 Distributionally robust optimization with unbalanced OT

In the KL divergence DRO described in the previous subsection, uncertainty set is limited to distributions with the same support as the empirical measure $\mu = \frac{1}{n}\sum_i \delta_{x_i}$. Another popular approach, Wasserstein DRO (WDRO) (Mohajerin Esfahani & Kuhn, 2018; Sinha et al., 2020), considers the worst-case risk over shifts within a Wasserstein (OT) ball around a reference measure μ instead of the KL-ball in (16), thus including continuous probability measures. Unfortunately, this approach is not resilient to outliers that are geometrically far from the clean distribution since OT metric is sensitive to them (Nietert et al., 2023). A natural generalization is to switch to semi-balanced OT (Liero et al., 2018; Chizat et al., 2019; Kondratyev et al., 2016), which replaces a hard constraint on one of the marginals with a mismatch penalty function, e.g.,

$$W_\beta(\nu, \mu) = \inf_{\substack{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{X}) \\ \pi_1 = \nu}} \int_{\mathcal{X}\times\mathcal{X}} c(x,z)\,\mathrm{d}\pi(x,z) + \beta\, D_{KL}(\pi_2, \mu),$$

where π_1 and π_2 are first and second marginals of π, respectively, $\beta > 0$ is the marginal penalty parameter. Intuitively, this discrepancy measure allows to ignore some points (e.g., outliers) by paying a small price for mismatch in marginals. The (penalty-form) DRO problem can be written as

$$\min_{\theta\in\Theta} \max_{\nu\in\mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} \ell(\theta, x)\,\mathrm{d}\nu(x) - \lambda W_\beta(\nu, \mu),$$

where $\lambda > 0$ is the Lagrangian penalty parameter. Using standard duality, Wang et al. (2024) showed that when $\mu = \frac{1}{n}\sum_i \delta_{x_i}$ is the empirical distribution, this is equivalent to

$$\min_{\theta\in\Theta} F(\theta) := \lambda\beta\, \log\Big(\frac{1}{n}\sum_{i=1}^n e^{\hat{\ell}_i(\theta)/(\lambda\beta)}\Big) \quad \text{with} \quad \hat{\ell}_i(\theta) := \sup_{z\in\mathcal{X}} \{\ell(\theta; z) - \lambda c(z, x_i)\}, \tag{21}$$

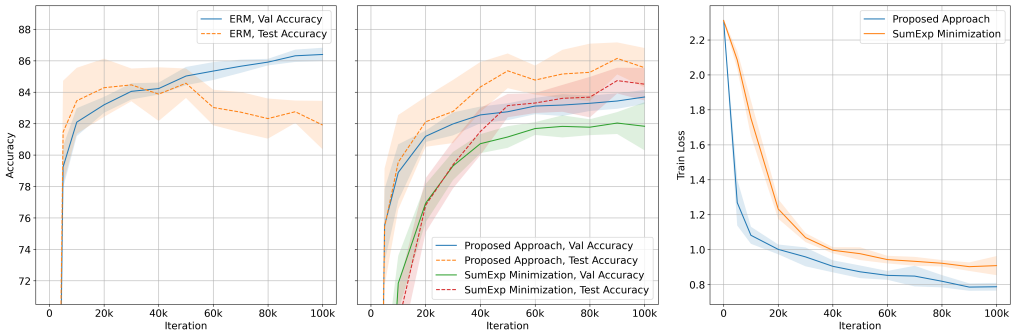

Figure 3: Performance of ERM and two DRO approaches on MNIST with noisy labels. Left: ERM accuracy on the noisy validation set vs. clean test set. Middle: validation vs. test accuracy for DRO approaches. Right: training loss $F(\theta)$ from (21).

To avoid the costly gradient computation of LogSumExp, the authors drop the logarithm and use SGD to optimize the sum of exponents,

$$\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} e^{\hat{\ell}_i(\theta)/(\lambda\beta)}. \tag{22}$$

The major downside of this approach is that the exponent terms have a large variance, and SGD is prone to floating-point exceptions (overflow) unless a very small stepsize is tuned, which slows down the convergence and can be time-consuming and unstable in practice.

Our approach. To overcome this issue, we propose leveraging the approximation (6), which leads to the problem

$$\min_{\substack{\theta\in\Theta\\ \alpha\in\mathbb{R}}} \frac{1}{n}\sum_{i=1}^{n}\Big\{\alpha + \frac{\lambda\beta}{\rho}\log\big(1+\rho e^{(\hat{\ell}_i(\theta)-\alpha)/(\lambda\beta)}\big)\Big\}, \tag{23}$$

where $\rho > 0$ is a parameter controlling the accuracy of the approximation. This approximation can be efficiently optimized with SGD. Note that our method can also be applied to other DRO algorithms such as Sinkhorn DRO (Wang et al., 2021), which we omit to avoid redundancy.

Experiments. We consider MNIST dataset (Deng, 2012) with train and validation labels corrupted by feature-dependent noise (see Algan & Ulusoy, 2020) (noise ratio 25%), and original (clean) test labels. Let θ denote weights of a CNN with two convolutional layers (32 and 64 channels, kernel size 3, ReLU activations, and 2×2 max pooling), followed by a fully connected classifier with one hidden layer of 128 units, and let $\ell(\theta; z)$ be its cross entropy loss on object z. In the experiment, SGD (with batch size 1) is applied to problems (22) (baseline) and (23) (proposed approach). We consider values of the stepsize $\eta \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$. Parameters λ and β were set to 1 since smaller values required a smaller stepsize and resulted in a slow convergence, while larger values pushed the model towards fitting the noisy distribution instead of the true one. Approximation accuracy parameter ρ in (23) was set to 0.1. For the inner maximization problem in (21), just 5 iterations of Nesterov's accelerated gradient method were sufficient to reach plateau in terms of the objective value. Additionally, we used SGD for the usual empirical risk minimization (ERM) to observe the effects of conventional (non-robust) training on noisy data.

Figure 3 demonstrates the performance of different approaches with best hyperparameter η (10^{-3} for ERM, 10^{-4} for the proposed approach, and 10^{-5} for the baseline). Shaded regions indicate $\pm$ one standard deviation across 10 runs, except that for the baseline approach (22) we excluded a single run that caused a floating-point exception. The plot on the left illustrates that ERM fits to the corrupted data well (accuracy on the noisy validation set is increasing) which results in decreasing accuracy on the clean test set. In contrast, the plot in the middle shows that an increase in validation accuracy results in the increase in test accuracy for both DRO approaches, which indicates that they are more capable at learning the underlying clean distribution. The plot on the right shows the train loss $F(\theta)$ from (21). As seen from the figure, the proposed approach converges faster than the baseline. This is caused by the fact that the baseline requires a small stepsize to avoid overflows.

4 CONCLUSION

We introduce a novel approximation to the log partition function (and in particular, to LogSumExp), which arises in numerous applications across machine learning and optimization. In the dual formulation, it corresponds to the safe KL divergence. Our LogSumExp approximation preserves convexity and smoothness, and can be efficiently minimized using stochastic gradient methods. Importantly, the respective gradient estimator has controllable bias independent of batch size, in contrast to prior approaches. Our empirical results highlight the practical advantages of the proposed approximation across tasks in continuous entropy-regularized OT and DRO. An important direction for future work is to leverage the approximation for other applications, where the LogSumExp function and duality of the KL divergence play a role.

REFERENCES

Görkem Algan and Ilkay Ulusoy. Label noise types and their effects on deep learning. *arXiv preprint arXiv:2003.10471*, 2020.

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. American Mathematical Soc., 2000. ISBN 978-0-8218-4302-4.

Arip Asadulaev, Alexander Korotin, Vage Egiazarian, Petr Mokrov, and Evgeny Burnaev. Neural optimal transport with general cost functionals. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gIiz7tBtYZ.

Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, February 2013. ISSN 0025-1909, 1526-5501. doi: 10.1287/mnsc.1120.1641. URL http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1120.1641.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Jeremiah Birrell, Paul Dupuis, Markos A. Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f,gamma)-divergences: Interpolating between f-divergences and integral probability metrics. *Journal of Machine Learning Research*, 23(39):1–70, 2022. URL http://jmlr.org/papers/v23/21-0100.html.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Pierre Blanchard, Desmond J. Higham, and Nicholas J. Higham. Accurately computing the log-sum-exp and softmax functions. *IMA Journal of Numerical Analysis*, 41(4):2311–2330, 2021. doi: 10.1093/imanum/draa038. URL https://academic.oup.com/imajna/article/41/4/2311/5893596.

Guillaume Bouchard. Efficient bounds for the softmax function and applications to approximate inference in hybrid models. In *NIPS 2007 workshop for approximate Bayesian inference in continuous/hybrid systems*, volume 6, 2007.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulation. *arXiv:1508.05216 [math]*, February 2019. URL http://arxiv.org/abs/1508.05216. arXiv: 1508.05216.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Mara Daniels, Tyler Maunu, and Paul Hand. Score-based generative neural networks for large-scale optimal transport. *Advances in neural information processing systems*, 34:12955–12965, 2021.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3440–3448. Curran Associates, Inc., 2016.

Josiah Willard Gibbs. *Elementary principles in statistical mechanics: developed with especial reference to the rational foundations of thermodynamics*. C. Scribner's sons, 1902.

David Goldberg. What every computer scientist should know about floating-point arithmetic. *ACM computing surveys (CSUR)*, 23(1):5–48, 1991.

Nikita Gushchin, Alexander Kolesov, Alexander Korotin, Dmitry P Vetrov, and Evgeny Burnaev. Entropic neural optimal transport via diffusion processes. *Advances in Neural Information Processing Systems*, 36:75517–75544, 2023.

Nicholas J. Higham. What is the log-sum-exp function? Blog post, *"What Is" series*, January 2021. URL https://nhigham.com/2021/01/05/what-is-the-log-sum-exp-function/. Available at nhigham.com.

Zhaolin Hu and L Jeff Hong. Kullback-Leibler Divergence Constrained Distributionally Robust Optimization.

Kelvin Kan, James G. Nagy, and Lars Ruthotto. Lsemink: A modified newton–krylov method for log-sum-exp minimization. *arXiv preprint arXiv:2307.04871*, 2023. URL https://arxiv.org/abs/2307.04871.

Leonid Kantorovich. On the translocation of masses. *(Doklady) Acad. Sci. URSS (N.S.)*, 37:199–201, 1942.

Mohammad Emtiyaz Khan and Didrik Nielsen. Fast yet simple natural-gradient descent for variational inference in complex models. In *2018 International Symposium on Information Theory and Its Applications (ISITA)*, pp. 31–35. IEEE, 2018.

Mohammad Emtiyaz Khan and Håvard Rue. The Bayesian learning rule. (arXiv:2107.04562), June 2023.

Stanislav Kondratyev, Léonard Monsaingeon, and Dmitry Vorotnikov. A new optimal transport distance on the space of finite Radon measures. *Advances in Differential Equations*, 21(11/12): 1117 – 1164, 2016. doi: 10.57262/ade/1476369298. URL https://doi.org/10.57262/ade/1476369298.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=d8CBRlWNkqH.

Alexander Korotin, Nikita Gushchin, and Evgeny Burnaev. Light schrödinger bridge. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WhZoCLRWYJ.

Daniel Kuhn, Soroosh Shafiee, and Wolfram Wiesemann. Distributionally Robust Optimization. (arXiv:2411.02549), November 2024. doi: 10.48550/arXiv.2411.02549.

Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in neural information processing systems*, 33:8847–8860, 2020.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, March 2018. ISSN 0020-9910, 1432-1297. doi: 10.1007/s00222-017-0759-8. URL http://link.springer.com/10.1007/s00222-017-0759-8.

Yucen Luo, Alex Beatson, Mohammad Norouzi, Jun Zhu, David Duvenaud, Ryan P. Adams, and Ricky T. Q. Chen. Sumo: Unbiased estimation of log marginal probability for latent variable models. *arXiv preprint arXiv:2004.00353*, 2020. URL https://arxiv.org/abs/2004.00353. Version v2, Apr 2020.

Anne-Marie Lyne, Mark Girolami, Yves Atchadé, Heiko Strathmann, and Daniel Simpson. On russian roulette estimates for bayesian inference with doubly-intractable likelihoods. *Statistical Science*, 30(4):443–467, 2015. doi: 10.1214/15-STS523. URL https://projecteuclid.org/journals/statistical-science/volume-30/issue-4/On-Russian-Roulette-Estimates-for-Bayesian-Inference-with-Doubly-Intractable/10.1214/15-STS523.full.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, September 2018. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-017-1172-1.

Petr Mokrov, Alexander Korotin, Alexander Kolesov, Nikita Gushchin, and Evgeny Burnaev. Energy-guided entropic neural optimal transport. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=d6tUsZeVs7.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer International Publishing, 2018.

Frank Nielsen and Ke Sun. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. *Entropy*, 18(12):442, 2016.

Sloan Nietert, Ziv Goldfeld, and Soroosh Shafiee. Outlier-robust wasserstein dro. *Advances in Neural Information Processing Systems*, 36:62792–62820, 2023.

R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33 (3):291–297, 1997.

EY Pee and Johannes O Royset. On solving large-scale finite minimax problems using exponential smoothing. *Journal of optimization theory and applications*, 148(2):390–421, 2011.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237. doi: 10.1561/2200000073. URL http://dx.doi.org/10.1561/2200000073. arXiv:1803.00567.

Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge university press, 2025.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Springer International Publishing, 2015. ISBN 9783319208282. doi: 10.1007/978-3-319-20828-2. URL http://dx.doi.org/10.1007/978-3-319-20828-2.

Vivien Seguy, Bharath Bhushan Damodaran, Remi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. In *ICLR 2018-International Conference on Learning Representations*, pp. 1–15, 2018.

Aras Selvi, Aharon Ben-Tal, Ruud Brekelmans, and Dick den Hertog. Convex maximization via adjustable robust optimization. Technical Report 7881, Optimization-Online, 2020. URL https://optimization-online.org/wp-content/uploads/2020/07/7881.pdf. Revised September 2, 2021.

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv:1710.10571 [cs, stat]*, May 2020.

Tasuku Soma and Yuichi Yoshida. Statistical learning with conditional value at risk. *arXiv preprint arXiv:2002.05826*, 2020.

Ryan Spring and Anshumali Shrivastava. A new unbiased and efficient class of lsh-based samplers and estimators for partition function computation in log-linear models. *arXiv preprint arXiv:1703.05160*, 2017. URL https://arxiv.org/abs/1703.05160. Mar 2017.

Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Alexey Naumov, Pierre Perrault, Michal Valko, and Pierre Menard. Demonstration-regularized RL. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=lF2aip4Scn.

Michalis K. Titsias. One-vs-each approximation to softmax for scalable estimation of probabilities. *Advances in Neural Information Processing Systems*, 29, 2016.

George Tucker, Andriy Mnih, Chris J. Maddison, Dieterich Lawson, and Jascha Sohl-Dickstein. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. *arXiv preprint arXiv:1703.07370*, 2017. URL https://arxiv.org/abs/1703.07370. v4, Mar 2017.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *arXiv preprint arXiv:2109.11926*, 2021.

Zifan Wang, Yi Shen, Michael Zavlanos, and Karl H Johansson. Outlier-robust distributionally robust optimization via unbalanced optimal transport. *Advances in Neural Information Processing Systems*, 37:52189–52214, 2024.

Xingyu Zhou. On the fenchel duality between strong convexity and lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*, 2018. URL https://arxiv.org/abs/1803.06573. Version v1, 17 Mar 2018.

A Proofs for Section 2

Proof of Proposition 2.4. (i,ii) Consider the function $g(t) := \frac{\ln(1+t)}{t}$. It is decreasing and convex on $(0,\infty)$, $g(t) \to 1$ and $g'(t) \to -\frac{1}{2}$ as $t \to 0+$. Note that

$$F_\rho(\varphi;\mu) = \inf_{\alpha\in\mathbb{R}} \alpha - 1 + \int_{\mathcal{X}} e^{\varphi(x)-\alpha} g\left(\rho e^{\varphi(x)-\alpha}\right) \mathrm{d}\mu(x).$$

Then (i) follows immediately from (6) and the monotonicity of g. The monotone convergence theorem yields (ii) since

$$F(\varphi;\mu) = \inf_{\alpha\in\mathbb{R}} \alpha - 1 + \int_{\mathcal{X}} e^{\varphi(x)-\alpha} \mathrm{d}\mu(x).$$

Now, let us prove (iii). Consider the optimal α_ρ, satisfying (7). By Jensen's inequality

$$\begin{aligned}\int_{\mathcal{X}} \ln\left(1+\rho e^{\varphi(x)-\alpha_\rho}\right) \mathrm{d}\mu(x) &= -\int_{\mathcal{X}} \ln\left(1 - \frac{\rho e^{\varphi(x)-\alpha_\rho}}{1+\rho e^{\varphi(x)-\alpha_\rho}}\right) \mathrm{d}\mu(x) \\ &\geq -\ln\left(1 - \int_{\mathcal{X}} \frac{\rho e^{\varphi(x)-\alpha_\rho}}{1+\rho e^{\varphi(x)-\alpha_\rho}} \mathrm{d}\mu(x)\right) = -\ln(1-\rho),\end{aligned}$$

thus

$$F_\rho(\varphi;\mu) = \alpha_\rho - 1 + \frac{1}{\rho}\int_{\mathcal{X}} \ln\left(1+\rho e^{\varphi(x)-\alpha_\rho}\right) \mathrm{d}\mu(x) \geq \alpha_\rho - 1 - \frac{\ln(1-\rho)}{\rho} \geq \alpha_\rho + \frac{\rho}{2}. \quad (24)$$

It remains to get a lower bound on α_ρ. By the monotonicity of $\frac{t}{1+t}$ we deduce that $\alpha_\rho \geq \alpha$ for any α such that

$$\int_{\mathcal{X}} \frac{\rho e^{\varphi(x)-\alpha}}{1+\rho e^{\varphi(x)-\alpha}}\,\mathrm{d}\mu(x) \geq \rho.$$

Since $\frac{t}{1+t} \geq t - t^2$,

$$\int_{\mathcal{X}} \frac{\rho e^{\varphi(x)-\alpha}}{1+\rho e^{\varphi(x)-\alpha}}\,\mathrm{d}\mu(x) \geq \int_{\mathcal{X}} \left(\rho e^{\varphi(x)-\alpha} - \rho^2 e^{2\varphi(x)-2\alpha}\right)\,\mathrm{d}\mu(x) = \rho e^{F(\varphi;\mu)-\alpha} - \rho^2 e^{F(2\varphi;\mu)-2\alpha}.$$

Denoting $u := e^{F(\varphi;\mu)-\alpha}$, it is enough to find u such that

$$u - au^2 \geq 1, \quad \text{where } a := \rho e^{F(2\varphi;\mu)-2F(\varphi;\mu)} \leq \frac{1}{4}.$$

Thus, taking

$$u := \frac{1}{2a}\left(1 - \sqrt{1-4a}\right) \leq 1 + 4a,$$

we obtain

$$\alpha_\rho \geq F(\varphi;\mu) - \ln u \geq F(\varphi;\mu) - \ln(1+4a) \geq F(\varphi;\mu) - 4a.$$

Combining this with (24), we get (8).

(iv) Finally, let $\varphi(x) \leq M$ for all $x \in \mathcal{X}$. Then by concavity

$$\int_{\mathcal{X}} \ln\left(1+\rho e^{\varphi(x)-\alpha}\right)\,\mathrm{d}\mu(x) \geq \int_{\mathcal{X}} e^{\varphi(x)-M}\ln\left(1+\rho e^{M-\alpha}\right)\,\mathrm{d}\mu(x) = e^{F(\varphi;\mu)-M}\ln\left(1+\rho e^{M-\alpha}\right)$$

for all $\alpha \in \mathbb{R}$. Therefore,

$$\begin{aligned} F_\rho(\varphi;\mu) &\geq \min_\alpha \alpha - 1 + \frac{e^{F(\varphi;\mu)-M}}{\rho}\ln\left(1+\rho e^{M-\alpha}\right) \\ &= F(\varphi;\mu) - 1 - \frac{1-\rho e^{M-F(\varphi;\mu)}}{\rho e^{M-F(\varphi;\mu)}}\ln\left(1-\rho e^{M-F(\varphi;\mu)}\right) \\ &\geq F(\varphi;\mu) - \rho e^{M-F(\varphi;\mu)}. \end{aligned}$$

Here we used the inequality

$$\frac{1-t}{t}\ln(1-t) \leq t-1, \quad 0<t<1.$$

□

Proof of Corollary 2.5. Set $\mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{a_i} \in \mathcal{P}(\mathbb{R})$. Then

$$\mathrm{LogSumExp}(a_1,\dots,a_n) = \ln n + \ln\left(\int_{\mathcal{X}} x\,\mathrm{d}\mu_n(x)\right) = \ln n + F(id;\mu_n)$$

and

$$\begin{aligned} \inf_{\alpha\in\mathbb{R}} \alpha - 1 + \frac{1}{\rho}\sum_{i=1}^n \ln(1+\rho e^{a_i-\alpha}) &= \inf_{\alpha\in\mathbb{R}} \alpha - 1 + \frac{n}{\rho}\int_{\mathcal{X}} \ln(1+\rho e^{x-\alpha})\,\mathrm{d}\mu_n(x) \\ &= \inf_{\alpha\in\mathbb{R}} \alpha - 1 + \frac{n}{\rho}\int_{\mathcal{X}} \ln\left(1+\frac{\rho}{n}e^{x-\alpha+\ln n}\right)\,\mathrm{d}\mu_n(x) \\ &= \ln n + F_{\rho/n}(id;\mu_n). \end{aligned}$$

Since

$$e^{F(id;\mu_n)-\max_i a_i} = \frac{\sum_{i=1}^n e^{a_i}}{n\max_i e^{a_i}} \geq \frac{1}{n} > \frac{\rho}{n},$$

Proposition 2.4(i,iv) yields

$$F(id;\mu_n) - \rho \leq F_{\rho/n}(id;\mu_n) \leq F(id;\mu_n).$$

The claim follows. □

Proof of Proposition 2.6. As $\lambda \ln(1 + e^{t/\lambda}) > t_+ := \max\{0, t\}$, we get

$$\begin{aligned}\lambda F_\rho(\varphi/\lambda; \mu) &\geq \lambda \inf_{\alpha \in \mathbb{R}} \alpha - 1 + \frac{1}{\rho} \int_{\mathcal{X}} \left(\ln \rho + \frac{\varphi(x)}{\lambda} - \alpha \right)_+ \mathrm{d}\mu(x) \\ &= \lambda(\ln \rho - 1) + \inf_{\alpha \in \mathbb{R}} \alpha + \frac{1}{\rho} \int_{\mathcal{X}} (\varphi(x) - \alpha)_+ \, \mathrm{d}\mu(x).\end{aligned}$$

The infimum in the r.h.s. is the variational formula for CVaR (9), thus we get the first inequality in (10). The second inequality can be obtained in a similar way using that $\lambda \ln(1 + e^{t/\lambda}) < t_+ + \lambda$. □

Proof of Lemma 2.7. Recall that we have

$$f_\rho(t) = \frac{1}{\rho}((\rho t)\log(\rho t) + (1 - \rho t)\log(1 - \rho t)) + 1 - t \log \rho.$$

Simplifying, we obtain

$$f_\rho(t) = t \log t + \frac{1}{\rho}(1 - \rho t)\log(1 - \rho t) + 1.$$

The first derivative is calculated as follows:

$$\frac{d}{dt}\big(t \log t\big) = \log t + 1, \qquad \frac{d}{dt}\left(\frac{1}{\rho}(1 - \rho t)\log(1 - \rho t)\right) = -\big(\log(1 - \rho t) + 1\big),$$

so

$$f'_\rho(t) = (\log t + 1) - \big(\log(1 - \rho t) + 1\big) = \log t - \log(1 - \rho t) = \log\left(\frac{t}{1 - \rho t}\right).$$

The second derivative is calculated as follows:

$$\frac{d}{dt}(\log t) = \frac{1}{t}, \qquad \frac{d}{dt}(\log(1 - \rho t)) = -\frac{\rho}{1 - \rho t},$$

thus

$$f''_\rho(t) = \frac{1}{t} + \frac{\rho}{1 - \rho t} = \frac{1}{t(1 - \rho t)}.$$

By symmetry, we can see that the minimum value of the second derivative is achieved at $t^* = \frac{1}{2\rho}$, and it is equal to 4ρ. Thus, for all $t \in \mathrm{dom} f_\rho$, we have that $f''_\rho(t) \geq 4\rho > \rho$. Thus, by (Nesterov, 2018, Theorem 2.1.11), f_ρ is ρ-strongly convex. By (Zhou, 2018, Theorem 1) this also implies that its conjugate function f^*_ρ is $\frac{1}{\rho}$-smooth. □

B Additional materials on entropic OT

B.1 Related works on eOT

This subsection provides an overview of selected works on continuous entropy-regularized optimal transport. In (Genevay et al., 2016), the authors tackled this problem by introducing an RKHS and optimizing the dual function (12) with SGD. This approach was extended by Seguy et al. (2018), who parameterized the dual potentials with neural networks instead of an RKHS to improve scalability. Subsequently, Daniels et al. (2021) leverage this approach to approximate the optimal transport plan, using it to develop a score-based generative model. Although this direction mostly results in computationally efficient methods that works with a general cost function, a key drawback is that small values of the regularization coefficient ε cause numerical instabilities due to the exponential term in the dual objective; see Remark 3.1. The work by (Korotin et al., 2023) studies a more general formulation known as *weak OT*. The authors formulate it as a maximin problem and develop a neural-network-based algorithm under the assumption of a quadratic cost, a restriction that is later relaxed in (Asadulaev et al., 2024). However, these methods are computationally intensive due to their adversarial training nature. The paper by Mokrov et al. (2024) approaches eOT from the perspective of energy-based models. Unfortunately, the resulting solver is computationally expensive

as it involves iterative Langevin dynamics . Another popular approach to eOT in recent years is via the Schrödinger bridge (SB), e.g., (Gushchin et al., 2023). While SB-based solvers are also often computationally intensive, a more cost-efficient solution has been proposed by (Korotin et al., 2024). However, it relies on the quadratic cost assumption and does not support general cost. We would also like to note that a promising direction for future work is leveraging our approach for minimizing the objective (8) in (Korotin et al., 2024) to further improve scalability.

B.2 Experiment with RKHS representation of dual potentials

As mentioned earlier, LSOT (Seguy et al., 2018) is inspired by the continuous eOT approach of Genevay et al. (2016). This work considers a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ defined on $\mathcal{X}$, with a kernel κ, and applies SGD to solve the dual problem. Such approach suffers from the same numerical instability as LSOT, see Remark 3.1. As an alternative, we again consider the approximation (15) of the semi-dual objective which can also be maximized by SGD. Although the variable α is, in general, a function of x, we empirically found that tuning a common scalar value $\alpha \in \mathbb{R}$ for all samples works well in the experiments described below.

Analytic form of SGD iterates for both objectives can be derived as follows. By the property of RKHS, if $u \in \mathcal{H}$, then $u(x) = \langle u, \kappa(\cdot, x)\rangle_{\mathcal{H}}$. Therefore, the derivatives of f_ε take the form

$$\nabla_u f_\varepsilon(x, y, u, v) = \kappa(\cdot, x) - \exp\left(\frac{u(x) + v(y) - c(x, y)}{\varepsilon}\right)\kappa(\cdot, x),$$
$$\nabla_v f_\varepsilon(x, y, u, v) = \kappa(\cdot, y) - \exp\left(\frac{u(x) + v(y) - c(x, y)}{\varepsilon}\right)\kappa(\cdot, y).$$

Consequently, SGD iterates for the dual objective (12) can be conveniently written as

$$(u_k, v_k) = (u_0, v_0) + \sum_{i=1}^{k} \beta_i \left(\kappa\left(\cdot, x_i\right), \kappa\left(\cdot, y_i\right)\right) \tag{25}$$

$$\text{with } \beta_i \coloneqq \frac{C}{\sqrt{i}}\Big(1 - e^{\frac{u_{i-1}(x_i) + v_{i-1}(y_i) - c(x_i, y_i)}{\varepsilon}}\Big), \tag{26}$$

where (x_i, y_i) are i.i.d. samples from $\mu \otimes \nu$, and $C > 0$ is the initial stepsize. Similarly, SGD iterates for (15) are computed as follows:

$$v_k = v_0 + \sum_{i=1}^{k} \tilde{\beta}_i \kappa\left(\cdot, y_i\right),$$
$$\alpha_k = \alpha_0 - \sum_{i=1}^{k} \tilde{\beta}_i \quad \text{with } \beta_i \coloneqq \frac{C}{\sqrt{i}}\Big(1 - \sigma_\rho\Big(\tfrac{u_{i-1}(x_i) + v_{i-1}(y_i) - c(x_i, y_i)}{\varepsilon}\Big)\Big),$$

where $\sigma_\rho(t) \coloneqq \frac{e^t}{1 + \rho e^t}$.

Experiments. Consider a setup analogous to the one described in Section 5 of Genevay et al. (2016). Specifically, μ is a 1D Gaussian, and ν is a mixture of two Gaussians (see Figure 4 for a plot of densities). Gaussian kernel $\kappa(x, x') = \exp\left(-\frac{\|x - x'\|^2}{\sigma^2}\right)$ with a bandwidth hyperparameter $\sigma^2 > 0$ is used. The regularization coefficient is set to $\varepsilon = 0.01$. We consider kernel SGD (25) applied to the dual objective as a *baseline* approach (Genevay et al., 2016). We compare it to the proposed approach, namely, kernel SGD applied to the approximate semi-dual problem (14). For details on how the optimality gap is estimated, see Appendix B.

When applying kernel SGD to the dual and approximate semi-dual formulations, we consider hyperparameters $\sigma^2 \in \{0.1, 1, 10\}$ (kernel bandwidth), $C \in \{10^{-4}, 10^{-3}, \dots, 10\}$ (stepsize parameter), and $\rho \in \{0.03, 0.1, 0.3\}$ (approximation accuracy). Double floating-point precision is used. In the experiment, the proposed approach works best with $\sigma^2 = 10$, and $C = 1$ for $\rho \in \{0.03, 0.1\}$, $C = 10$ for $\rho = 0.3$. Baseline works best with $\sigma^2 \in \{0.1, 1\}$ and $C = 10^{-3}$. Figure 5 (left) shows performance of the two approaches. For clarity, we provide a zoomed-in view of the curves generated by the baseline in the middle. As seen from the figures, the baseline is extremely slow, which

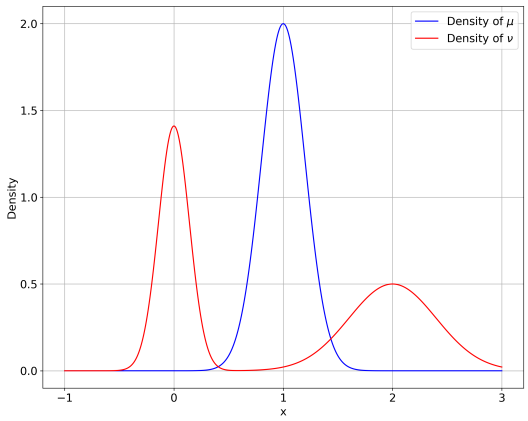

Figure 4: Densities of source and target distributions in the eOT experiment.

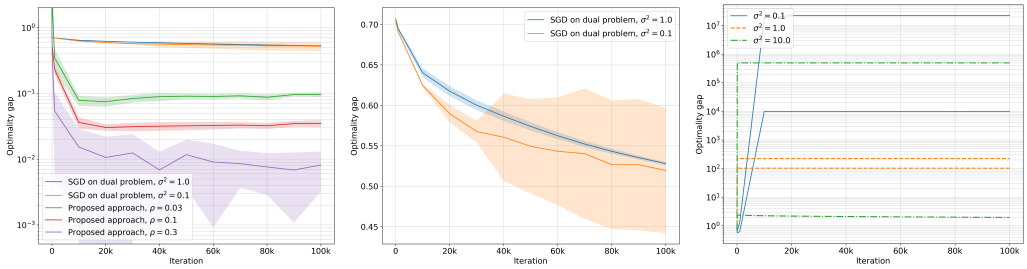

Figure 5: Left: convergence of kernel SGD applied to the dual objective (12) (blue and orange) and approximate semi-dual problem (14) (green, red and purple). Solid lines show average optimality gap across 20 runs, shaded regions indicate $\pm$ one standard deviation. Y-axis uses logarithmic scale. Middle: a zoomed-in view of blue and orange curves from the plot on the left. Right: examples of divergent optimality gap curves obtained by running the baseline approach with the stepsize parameter $C = 10^{-2}$.

happens due to the small stepsize. Larger values of C lead to numerical instabilities as illustrated by the plot on the right. Apparently, exponent causes a large magnitude of the gradient at a certain step, which brings an iterate to a region where it stagnates. On the contrary, our approximate semi-dual formulation permits larger stepsizes, which results in faster convergence. Indeed, the method usually achieves a relatively low optimality gap in about $2 \cdot 10^4$ iterations, and plateaus after that.

B.3 Computing a proxy for optimality gap

Optimality gap in the experiment is estimated as follows:

1. Test sets $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$ of size $N = 10^4$ are sampled from μ and ν. The corresponding empirical distributions are denoted $\hat{\mu}$ and $\hat{\nu}$, respectively.
2. Similarly to Genevay et al. (2016), we obtain a proxy $\hat{W}$ for $W(\mu, \nu)$ by solving the semi-discrete eOT problem

$$\max_{\mathbf{v} \in \mathbb{R}^N} \mathbb{E}_{X \sim \mu} \hat{h}_\varepsilon(X, \mathbf{v})$$

$$\text{with } \hat{h}_\varepsilon(x, \mathbf{v}) := \frac{1}{N} \sum_{i=1}^N v_i - \varepsilon \log\Big(\frac{1}{N} \sum_{i=1}^N e^{\frac{v_i - c(x, y_i)}{\varepsilon}} \Big) - \varepsilon,$$

which corresponds to replacing the expectation $\mathbb{E}_{Y \sim \nu}$ in (13) with the average over the test set $\mathbb{E}_{Y \sim \hat{\nu}}$. We preform 10 runs of SGD, each consisting of $2 \cdot 10^5$ iterations, and define $\hat{W}$ as largest achieved value on the test set, i.e., the largest $\mathbb{E}_{X \sim \hat{\mu}} \hat{h}_\varepsilon(X, \mathbf{v})$.

3. Finally, given a potential $v \in \mathcal{C}(\mathcal{X})$, we estimate the optimality gap as $\hat{W} - \mathbb{E}_{X\sim\hat{\mu}}\hat{h}_\varepsilon(X, \mathbf{v})$, where $\mathbf{v} = (v(y_1), \dots, v(y_N))^\top$ is the evaluation of v on the test set.

C Properties of SoftPlus

Let $F(x) = \log(1 + e^{f(x)})$, then

$$\nabla F(x) = \sigma(f(x))\nabla f(x), \tag{27}$$

$$\nabla^2 F(x) = \sigma(f(x))\nabla^2 f(x) + \sigma(f(x))\big(1 - \sigma(f(x))\big)\nabla f(x)\nabla f(x)^\top. \tag{28}$$

Suppose $f(x)$ is L-smooth (possibly non-convex). Let us derive smoothness constant of F. We will use the following

Lemma C.1. *Consider function $f_a(x) = \sigma(x) + 2\sigma'(x)(x - a)$, $x \geq a$ with parameter $a \leq 0$. It holds $f_a(x) \leq 2 - \frac{a}{2}$.*

Proof. By the properties of the sigmoid function $\sigma(x)$, $\sigma'(x) \leq \frac{1}{4}$ and $\sigma(x) \leq 1$. Therefore, $f_a(x) \leq 1 + \frac{x-a}{2}$. If $x \leq 2$, the result follows. Let us now show that the derivative

$$\frac{d}{dx} f_a(x) = \sigma'(x)[3 + 2(1 - 2\sigma(x))(x - a)]$$

is negative if $x > 2$. Indeed, due to monotonicity of the sigmoid function $\sigma(x)$,

$$\sigma(x) > \sigma(2) > 0.88 \Rightarrow 2(1 - 2\sigma(x)) < -\frac{3}{2}.$$

Moreover, $x - a > 2$, so $3 + 2(1 - 2\sigma(x))(x - a) < 0$ and $\frac{d}{dx} f_a(x) < 0$. Therefore, if $x > 2$, then $f_a(x) < f_a(2) \leq 2 - \frac{a}{2}$. □

Proposition C.2. *Let $f \in C^1(\mathbb{R}^d)$ be L-smooth and bounded from below by $f_* \in \mathbb{R}$, then $F(x) = \log(1 + e^{f(x)})$ is smooth with parameter*

$$\begin{cases} \frac{4}{3}L & \text{if } f_* \geq 0, \\ \left(\frac{4}{3} - \frac{f_*}{2}\right) L & \text{if } f_* < 0. \end{cases} \tag{29}$$

Proof. W.l.o.g., we can assume that $f \in C^2$. From (28) and Lemma C.1 we get

$$\begin{aligned} \|\nabla^2 F(x)\| &\leq \sigma(f(x))\|\nabla^2 f(x)\| + \sigma'(f(x))\|\nabla f(x)\|^2 \\ &\leq L\sigma(f(x)) + 2L\sigma'(f(x))(f(x) - f_*) \\ &= L\big(\sigma(f(x)) + 2\sigma'(f(x))f(x)\big) - 2L\sigma'(f(x))f_*. \end{aligned}$$

Analyzing the function $h(t) \coloneqq \big(\sigma(t) + 2t\sigma'(t)\big)$, one can show that $\max_t h(t) < \frac{4}{3}$. Thus, in the case $f_* \geq 0$, using the fact that $\sigma'(t) > 0$ we obtain

$$\|\nabla^2 F(x)\| \leq Lh(f(x)) \leq \frac{4}{3}L.$$

Now, consider the case $f_* < 0$. Since $\sigma'(t) = \sigma(t)(1 - \sigma(t)) \leq \frac{1}{4}$,

$$\|\nabla^2 F(x)\| \leq Lh(f(x)) - 2L\sigma'(f(x))f_* \leq \frac{4}{3}L - \frac{L}{2}f_*.$$

The claim follows.

□

Remark C.3. *The factor $\frac{1}{2}$ in front of $-f_*$ in (29) can't be improved. Indeed, consider $f(x) = \frac{1}{2}(x - a)^2 - \frac{1}{2}a^2$ with $f_* = -\frac{1}{2}a^2$. The second derivative of $F(x) = \log(1 + e^{f(x)})$ is*

$$F''(x) = \sigma(f(x)) + \sigma(f(x))\,(1 - \sigma(f(x)))\,(x - a)^2,$$

$$F''(0) = \sigma(0) + \sigma(0)\,(1 - \sigma(0))\,a^2 = \frac{1}{2} + \frac{a^2}{4} = \frac{1}{2} - \frac{f_*}{2}.$$

Proposition C.4. *If f is convex, then $F(x) = \log(1 + e^{f(x)})$ is also convex.*

Proof. Trivially follows from (28). □

D LLM USAGE DISCLOSURE

In the preparation of this manuscript, large language models (LLMs) were used to improve the readability. All substantive contributions are solely by the authors.