Improved Stochastic Optimization of LogSumExp

Egor Gladin

Научная конференция ФКН

29.10.2025

The talk is based on the following work:

Egor Gladin, Alexey Kroshnin, Jia-Jie Zhu, Pavel Dvurechensky *Improved Stochastic Optimization of LogSumExp.* https://arxiv.org/abs/2509.24894

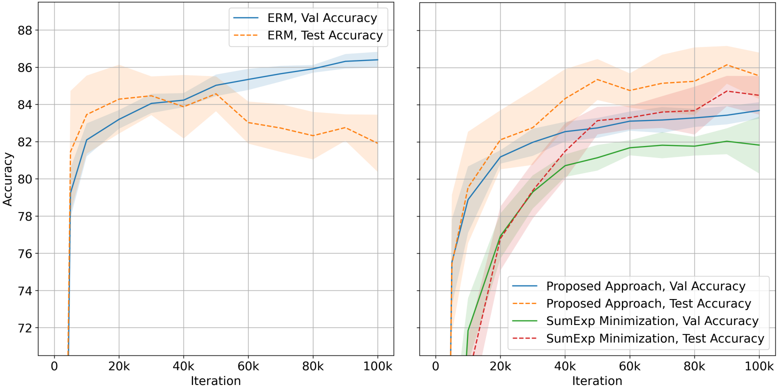
1. LogSumExp minimization — motivating examples;
2. SGD-friendly LogSumExp approximation;
3. Numerical experiments.

DRO

Distributionally Robust Optimization (DRO) aims to train an ML model robust to data distribution shifts.

Example: training a classifier on noisy MNIST.

- train and validation labels are corrupted, test labels are clean;
- left: ERM fits to the corrupted data well but fails on test;
- right: DRO approaches can learn the underlying clean distribution better.



A popular formulation in DRO is

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

where

- $\theta \in \Theta$ is the model parameters;
- $\ell_i(\theta)$ is the loss on the i-th example;
- Δ^n is the unit simplex in $\mathbb{R}^n$;
- $\hat{p} \in \Delta^n$ defines the empirical distribution (typically $\hat{p} = \frac{1}{n}\mathbf{1}$);
- D_{KL} is the Kullback–Leibler divergence defined as

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

which discourages distributions that are too far from the empirical one.

Analytic formula for maximization w.r.t. p leads to the problem

$$\min_{\theta \in \Theta} \mathcal{L}(\theta) := \lambda \log\left(\frac{1}{n} \sum_{i=1}^{n} e^{\ell_i(\theta)/\lambda}\right). \tag{2}$$

$$\min_{\theta\in\Theta} \mathcal{L}(\theta) := \lambda \log\Big(\frac{1}{n}\sum_{i=1}^{n} e^{\ell_i(\theta)/\lambda}\Big). \tag{3}$$

Challenge: When n is large, computing the full gradient $\nabla\mathcal{L}(\theta) = \sum_{i=1}^{n} p_i^*(\theta)\nabla\ell_i(\theta)$ with $p_i^*(\theta) := \frac{e^{\ell_i(\theta)/\lambda}}{\sum_j e^{\ell_j(\theta)/\lambda}}$ becomes costly.

Existing approach:

- sample a batch D;
- compute the respective softmax weights $p_i^D(\theta) := \frac{e^{\ell_i(\theta)/\lambda}}{\sum_{j\in D} e^{\ell_j(\theta)/\lambda}}$;
- define a gradient estimator by

$$\tilde{\nabla}_D\mathcal{L}(\theta) = \sum_{i\in D} p_i^D(\theta)\nabla\ell_i(\theta). \tag{4}$$

Problem: This introduces a bias and requires using large batch sizes to keep it sufficiently small. *A better approach to LogSumExp optimization is needed.*

LogSumExp optimization also arises in:

- entropy-regularized optimal transport (OT);
- minimax problems;
- multiclass classification with softmax probabilities;
- entropy-regularized reinforcement learning (RL);
- and many other applications...

Generalization of LogSumExp is the log-partition functional (a.k.a. free energy)

$$F(\varphi;\mu) := \ln \int e^{\varphi(x)} \mathrm{d}\mu(x), \tag{5}$$

mapping a measurable function φ to $(-\infty, \infty]$ based on a probability measure μ.

Theorem

Let $0 < \rho < 1$. The function

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

satisfies

$$F_\rho(\varphi;\mu) - O(\rho) \le F(\varphi;\mu) \le F_\rho(\varphi;\mu)$$

for any probability measure μ and measurable function φ.

In applications, φ is often defined as the parametric loss function $L(x, \theta)$:

$$\min_{\theta \in \Theta} F(\theta) := \ln \int e^{L(x,\theta)} \mathrm{d}\mu(x).$$

Relaxed problem

$$\min_{\theta \in \Theta, \alpha \in \mathbb{R}} G_\rho(\theta, \alpha) := \alpha + \log \rho - 1 + \frac{1}{\rho} \int \log\left(1 + e^{L(x,\theta) - \alpha}\right) \mathrm{d}\mu(x). \tag{7}$$

Denote $\sigma(t) := \frac{1}{1+e^{-t}}$, then we can define an unbiased gradient estimator

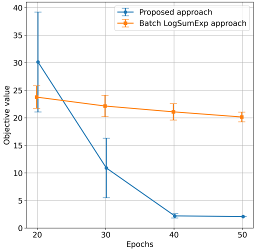
$$\nabla_\theta G_\rho(\theta, \alpha; x) = \frac{1}{\rho} \sigma\left(L(x, \theta) - \alpha\right) \nabla_\theta L(x, \theta),$$
$$\partial_\alpha G_\rho(\theta, \alpha; x) = 1 - \frac{1}{\rho} \sigma\left(L(x, \theta) - \alpha\right).$$

$$\min_{\theta\in\Theta,\alpha\in\mathbb{R}} G_\rho(\theta,\alpha) := \alpha + \log\rho - 1 + \frac{1}{\rho}\int \log\left(1 + e^{L(x,\theta)-\alpha}\right) \mathrm{d}\mu(x). \tag{8}$$

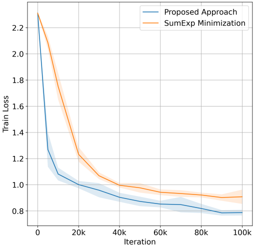
Properties of the objective:

- G_ρ is convex in α;
- if L is convex in θ, then G_ρ is jointly convex;
- if L is bounded from below, then G_ρ is Lipschitz-smooth on $\Theta \times (-\infty, a]$ for any $a \in \mathbb{R}$.

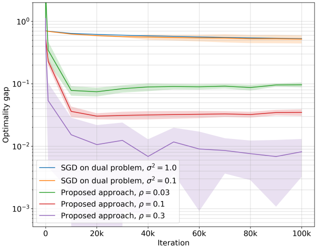
The proposed approach outperforms baselines in a number of experiments.



(a) DRO for regression



(b) DRO for classification



(c) Regularized optimal transport

Thank you for your attention!

Questions are welcome!

Contact: tg egorgladin