Adaptive Sampling for Risk-Averse Stochastic Learning

Sebastian Curi, Kfir Y. Levy,
Stefanie Jegelka, Andreas Krause

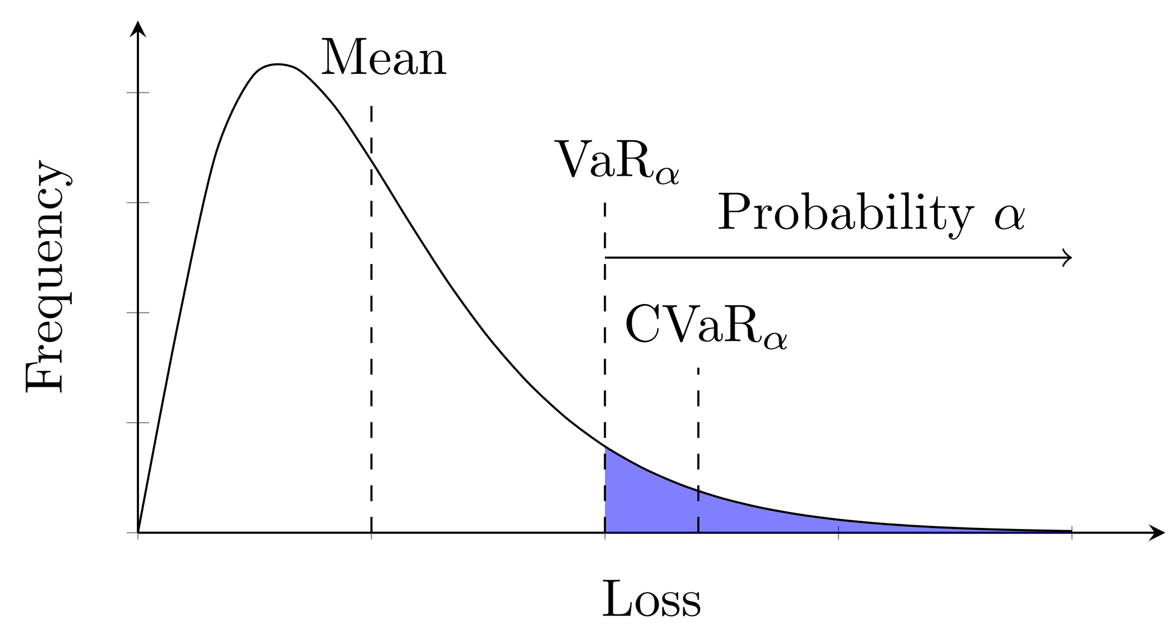
tldr: **AdaCVaR**, a novel algorithm for CVaR optimization in deep learning

Paper

Code

What is the CVaR?

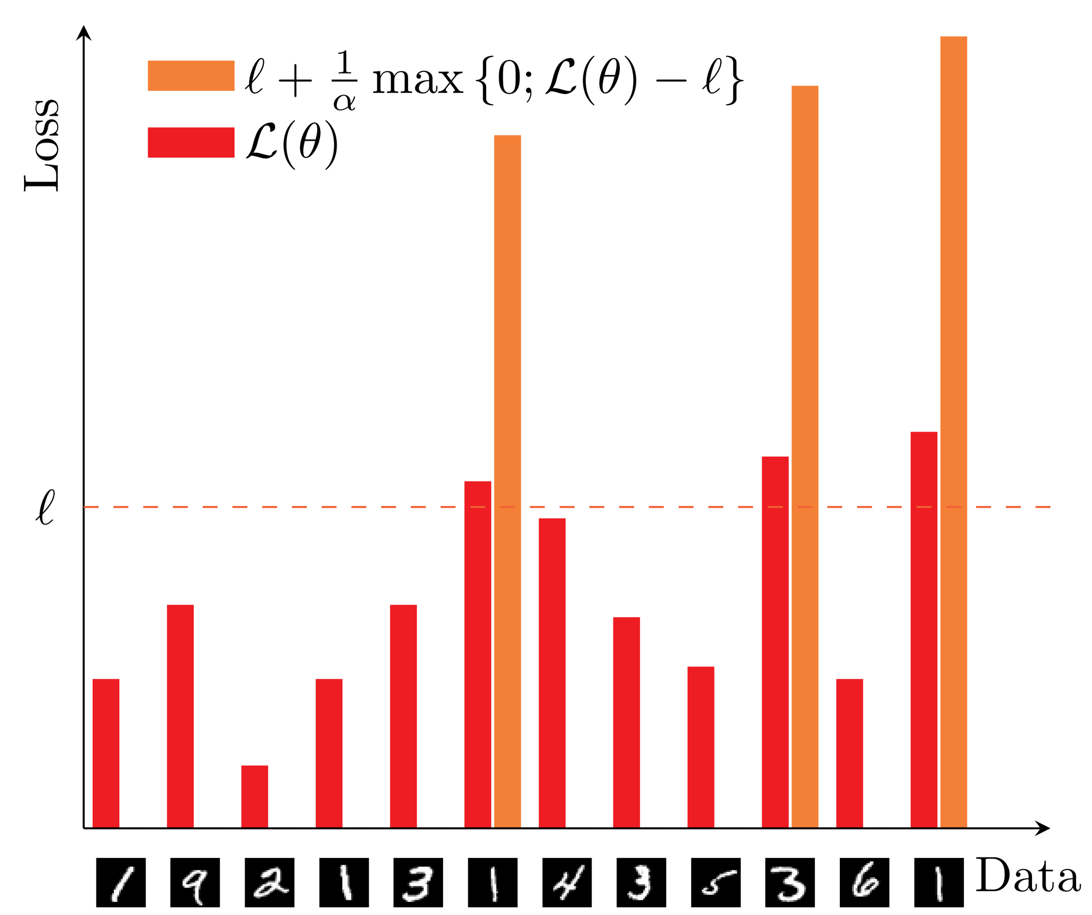
- In high-stake applications, we want to do well even in **rare** events.
- Standard **ERM** may *sacrifice large-but-rare* losses for the sake of performing well in average.
- Rather than focusing on the mean, the **CVaR** optimizes the average of the **tail** of the distribution and focuses on harder examples.



Related Work and Stochastic Optimization

Most of the previous work (e.g., Fan et al. (2019)) optimize the CVaR using the variational formula of Rockafellar & Uryasev (2000).

$$\min_\theta \mathrm{CVaR}_\alpha[\mathcal{L}(\theta)] = \min_{\theta, \ell \in \mathbb{R}} \ell + \frac{1}{\alpha}\mathbb{E}\left[\max\{0, \mathcal{L}(\theta) - \ell\}\right]$$



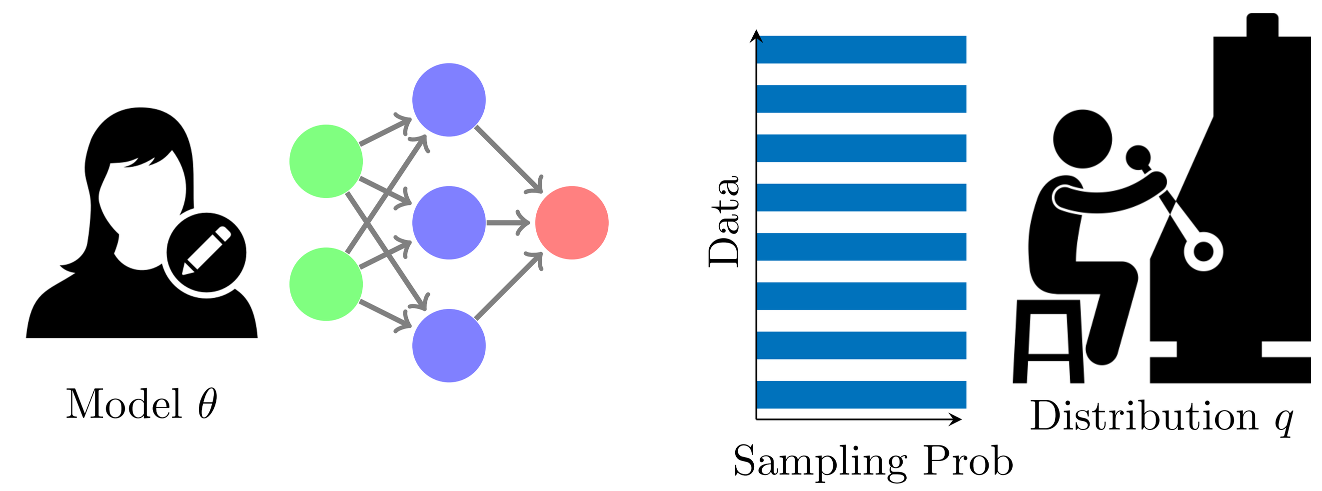
Unfortunately, this formula is not well suited for large-scale stochastic optimization. The **variance** of gradients is increased due to:

- Truncating the losses to zero
- Multiplying losses by $\frac{1}{\alpha}$

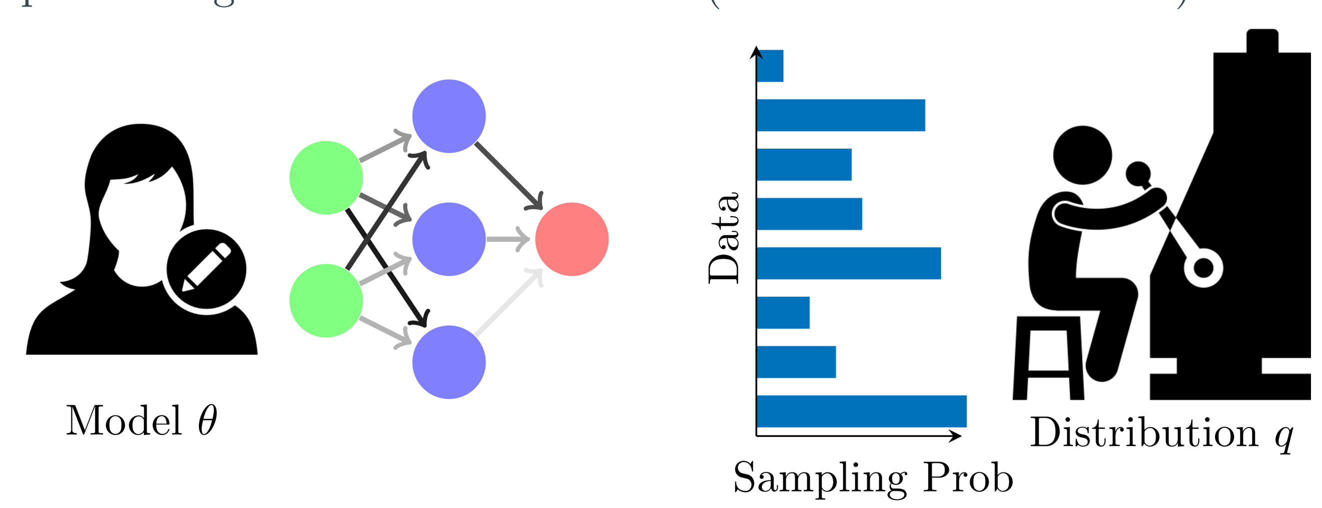
AdaCVaR: A DRO Game

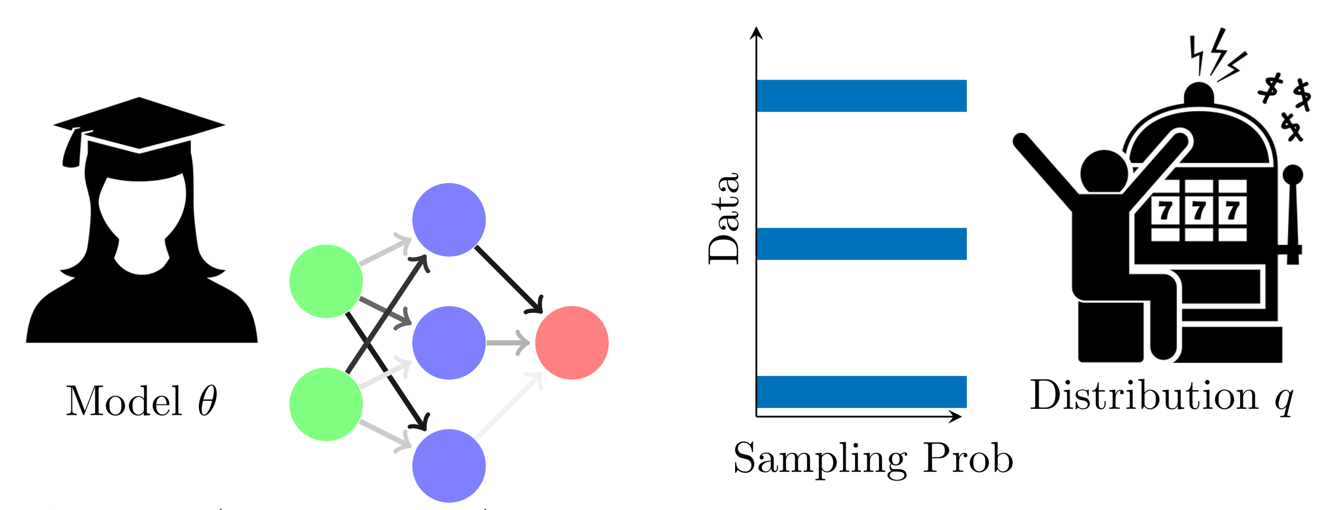
Instead of using the variational formula of Rockafellar & Uryasev (2000), we use the distirbutionally robust formulation of the CVaR (Shapiro et al. 2014).

$$\min_\theta \mathrm{CVaR}_\alpha[\mathcal{L}(\theta)] = \min_{\theta \in \Theta} \max_{q \in \mathcal{Q}^\alpha} \mathbb{E}_q[\mathcal{L}(\theta)]$$



- Game between a learner and a sampler. **Challenge**: DRO set is combinatorial.
- Sampler plays k.EXP3 from Alatur et al. (2020) to find the hardest distributions for the models the learner selects, ***adaptively***.
- Learner plays **SGD** on the examples proposed by the sampler.
- We exploit the problem structure i.e., combinatorial set with additive losses implementing k.EXP3 with **k-DPPs** (Kulesza & Taskar 2012).





Definition (Game Regret):

$$\mathrm{GameRegret}_T := \sum_{t=1}^T \mathbb{E}_{q^\star}\left[L(\theta_t)\right] - \mathbb{E}_{q_t}\left[L(\theta^\star)\right]$$

Theorem (AdaCVaR Regret):

$$\mathrm{GameRegret}_T = O(\sqrt{TN\log N} + \epsilon_{\mathrm{SGD}} T)$$

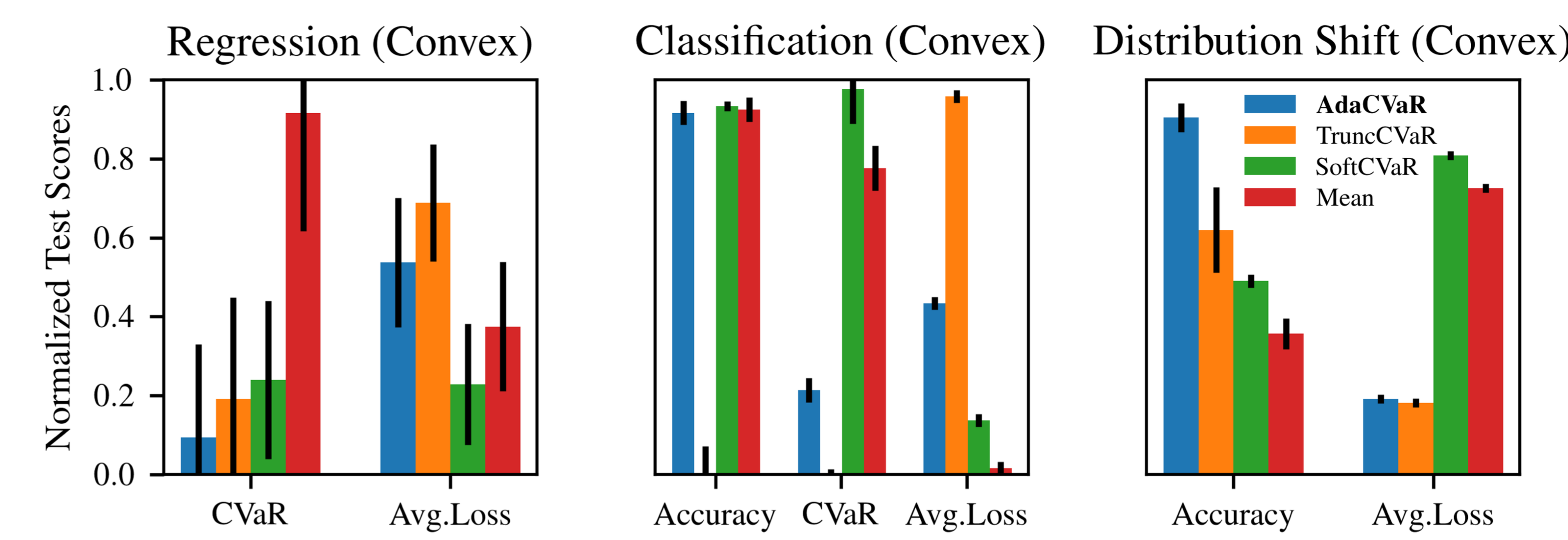
Non-convex: SGD error on ERM
Convex: $O\left(\frac{1}{\sqrt{T}}\right)$

Corollary (Online-to-Batch + Population Guarantee):

$$\mathbb{E}\mathrm{CVaR}_\alpha(\bar{\theta}) = O\left(\sqrt{\frac{N\log N}{T}} + \epsilon_{\mathrm{SGD}} + \frac{2}{\alpha}\sqrt{\frac{\log(2|\Theta|/\delta)}{N}}.\right)$$

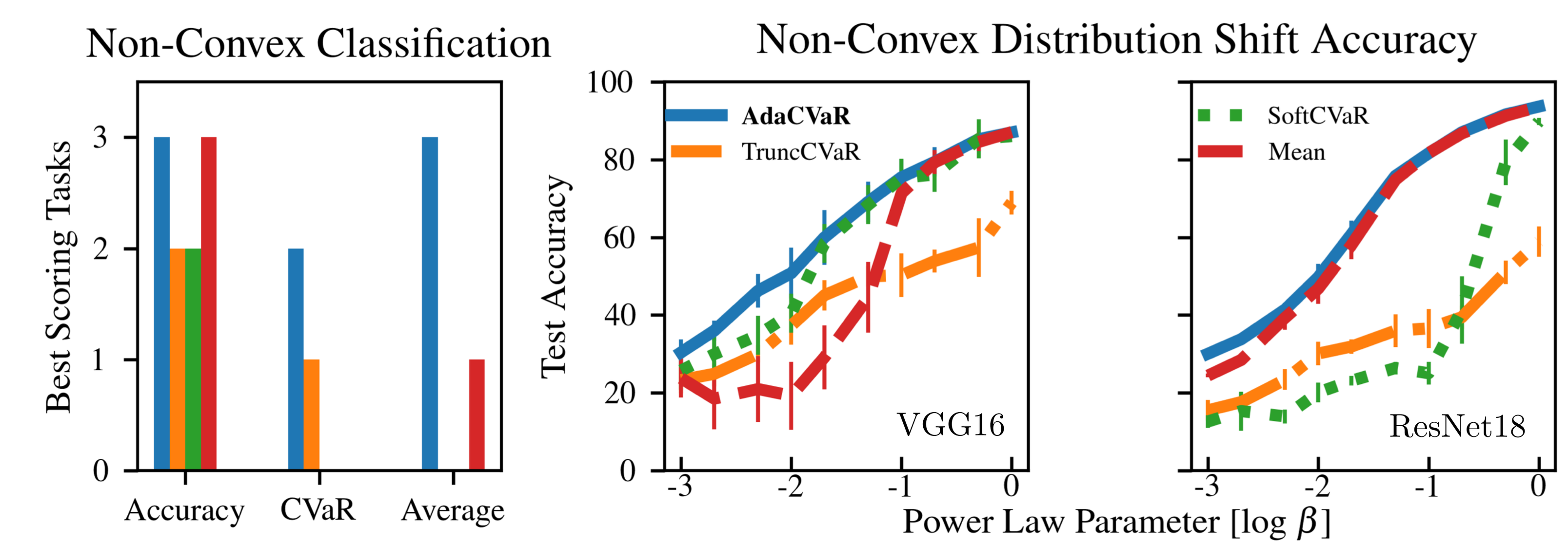
Experimental Results

Convex Optimization Tasks:



- AdaCVaR has lower CVaR in Regression.
- AdaCVaR has highest accuracy and low CVaR in Classification.
- AdaCVaR has highest accuracy and lowest CVaR with distribution shift.

Non-Convex Optimization Tasks:



- AdaCVaR has highest accuracy and lowest CVaR in image recognition.
- AdaCVaR performs consistently better under distribution shift.

References

Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. Journal of risk, 2, 21-42.

Shapiro, A., Dentcheva, D., & Ruszczyński, A. (2014). Lectures on stochastic programming: modeling and theory. Society for Industrial and Applied Mathematics.

Fan, Y., Lyu, S., Ying, Y., & Hu, B. (2017). Learning with average top-k loss. In Advances in neural information processing systems.

Alatur, P., Levy, K. Y., & Krause, A. (2020). Multi-player bandits: The adversarial case. Journal of Machine Learning Research, 21.

Kulesza, A., & Taskar, B. (2012). Determinantal Point Processes for Machine Learning. Foundations and Trends® in Machine Learning, 5(2–3), 123-286.