Black-Box Uniform Stability for Non-Euclidean Empirical Risk Minimization

Simon Vary[1] David Martínez Rubio[2] Patrick Rebeschini[1]

[1]Dept. of Statistic, University of Oxford [2]Signal Theory and Communications Dept., Carlos III University of Madrid

What is this poster about?

An black-box reduction method that turns a first-order optimization algorithm for **smooth convex** losses w.r.t. p-norms into a **uniformly stable** learning algorithm.

- Achieves optimal statistical risk bound
- Regularity of the loss w.r.t. non-Euclidean ℓ_p-norm
- Improves over ℓ_2-regularization in high-dimensional setting
- Generalization (via stability) of uniformly convex regularizers

Learning in Non-Euclidean Geometry

Setup: Loss $\ell : \mathcal{B}_p(R) \times \mathcal{Z} \to \mathbb{R}$ convex smooth w.r.t. $\|\cdot\|_p$, where $\mathcal{B}_p(R) \coloneqq \{x \in \mathbb{R}^d : \|x\|_p \leq R\}$, distribution $P \in \mathcal{P}$ supported on $\mathcal{Z}$, find

$$\tilde{x} \in \arg\min_{x \in \mathcal{B}_p(R)} f(x) \coloneqq \mathbb{E}_{z \sim P}\left[\ell(x, z)\right],$$

given *i.i.d.* samples $S = \{z_i\}_{i=1}^N \in \mathcal{Z}$ from P, empirical risk $f_S(x) = \frac{1}{N}\sum_{i=1}^N \ell(x, z_i)$.

Decomposition of the expected excess risk of estimator $\hat{x} = \mathcal{A}(S)$:

$$\begin{aligned}\mathbb{E}_S[f(\hat{x}) - f(\tilde{x})] &= \mathbb{E}_S[f(\hat{x}) - f_S(\hat{x})] + \mathbb{E}_S[f_S(\hat{x}) - f_S(\tilde{x})] \\ &\leq \underbrace{\mathbb{E}_S[f(\hat{x}) - f_S(\hat{x})]}_{\text{Statistics}} + \underbrace{\mathbb{E}_S[f_S(\hat{x}) - f_S(x^*)]}_{\text{Optimization}},\end{aligned}$$

using $\mathbb{E}_S[f_S(x)] = f(x)$ and $x^* = \arg\min_{x \in \mathcal{B}_p(R)} f_S(x)$.

Open Problem

Question: Given an optimization alg. $(x_t)_t$ that achieves optimal time complexity

$$f(x_t) - \inf_{x \in \mathcal{B}} f(x) \sim \inf_{\substack{(x_t)_t \\ \text{first-order solvers}}} \sup_f [f(x_t) - \inf_{x \in \mathcal{B}} f(x)],$$

design a statistical alg. $(x'_t)_t$ that, using S, achieves optimal statistical complexity

$$\mathbb{E}_S[f(x'_t) - \inf_{x \in \mathcal{B}} f(x)] \sim \inf_{\substack{(x'_t)_t \\ \text{first-order solvers}}} \sup_{P, \ell} \mathbb{E}_S[f(x'_t) - \inf_{x \in \mathcal{B}} f(x)].$$

Posed by Attia & Koren (2022), who solved for $p = 2$.

Main Idea

Apply an optimization algorithm to the ERM with added regularization:

$$x^*_\mu \in \arg\min_{x \in \mathcal{X}} f_S^{(\mu)}(x) \coloneqq f_S(x) + \mu\frac{\alpha}{p}\|x - x_0\|_p^p,$$

where $\alpha > 0$ ensures that $\psi(x) \coloneqq \frac{\alpha}{p}\|x - x_0\|_p^p$ is

- $(1, p)$-uniformly convex for $p \geq 2$, and
- $(1, p)$-Hölder smooth for $p \in (1, 2)$ w.r.t. ℓ_p-norm.

Also, $\psi(x)$ is locally smooth for $p \geq 2$ and strongly convex for $p \in (1, 2)$ w.r.t. ℓ_p-norm.

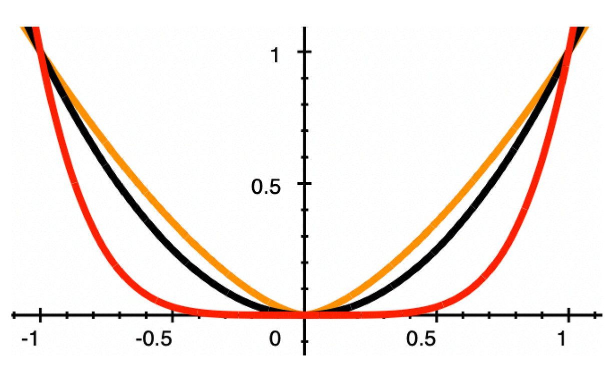
Uniformly Convex Regularization

(μ, ν)**-Uniform Convexity** of $\psi(x)$, $\nu \geq 2$:

$$\psi\left(tx + (1-t)y\right) \geq t\psi(x) + (1-t)\psi(y) + t(1-t)\frac{\mu}{\nu}\|x - y\|^\nu$$

(L, ν)**-Hölder smoothness** of $f(x)$, $\nu \in [1, 2]$:

$$f(y) \leq f(x) + \langle \nabla \ell(x), y - x\rangle + \frac{L}{\nu}\|x - y\|^\nu$$



References

1. Attia and Koren. *Uniform Stability for First-Order Empirical Risk Minimization*.ICLR, 2021.
2. Bousquet and Elisseeff. *Stability and Generalization.* Journal of Machine Learning Research, 2002.
3. Sridharan. *Learning From An Optimization Viewpoint.* PhD Thesis. Toyota Technological Institute at Chicago , 2012.
4. Levy & Duchi. *Necessary and Sufficient Geometries for Gradient Methods.* NeurIPS, 2019.

Algorithmic Stability after Uniformly Convex Regularization

Uniform Algorithmic Stability (Bousquet and Elisseeff; 2002):
Let $x' = \mathcal{A}(S_i)$ be the output of alg. $\mathcal{A}$ trained on $S_i = \{z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_N\}$. Then:

$$\underbrace{\mathbb{E}_S[f(\hat{x}) - f_S(\hat{x})]}_{\text{Statistics}} = \frac{1}{N}\sum_{i=1}^N \mathbb{E}[\ell(\hat{x}, z'_i) - \ell(x', z'_i)] \leq \sup_{z \in \mathcal{Z}} |\ell(\hat{x}; z) - \ell(x'; z)| \eqqcolon \varepsilon_{\text{stab}}(\mathcal{A}),$$

and we say the algorithm $\mathcal{A}(\cdot)$ is $\varepsilon_{\text{stab}}(\mathcal{A})$-uniformly stable.

Lemma (Stability after Uniformly Convex Regularization):
Let loss $\ell(\cdot, z)$ be convex G-Lipschitz w.r.t $\|\cdot\|$, $\mu\psi(x)$ is (μ, ν)-uniformly convex w.r.t $\|\cdot\|$. Then

$$\hat{x} \in_\varepsilon \arg\min_{x \in \mathcal{B}_{\|\cdot\|}(R)} \left[f_S(x) + \mu\psi(x)\right],$$

has its stability and optimization bounded as

$$\begin{aligned}\varepsilon_{\text{stab}}(\mathcal{A}) &\leq 3\left(\frac{2\nu}{n\mu}G^\nu\right)^{\frac{1}{\nu-1}} \\ \varepsilon_{\text{opt}}(\mathcal{A}) &\coloneqq f_S(\hat{x}) - \min_{x \in \mathcal{B}_p(R)} f_S(x) \leq 2\mu R^\nu,\end{aligned}$$

provided $\varepsilon \leq \min\{\mu R^\nu, (\nu/\mu)^{1/(\nu-1)}(2G/n)^{\nu/(\nu-1)}\}$.

Black-box Non-Euclidean Uniform Stability

Black-box scheme (with soft-restarts) that computes ε_i-approx. locally-strongly ($p \in (1, 2)$) / globally-uniformly ($p \geq 2$) convex regularized ERMs.

Optimization algorithm $\mathcal{A}(f_S^{(\mu)}, x_0, R, \hat{\varepsilon})$:

- for convex Hölder smooth functions,
- takes x_0, $R \geq 0$, and target accuracy $\hat{\varepsilon}$,

and outputs a point $\hat{x} \in \mathcal{B}_p(x_0, R)$ such that

$$f_S^{(\mu)}(\hat{x}) - \inf_{x \in \mathcal{B}_p(x_0, R)} f_S^{(\mu)}(x) \leq \hat{\varepsilon},$$

using at most $\hat{T}$ gradient oracle calls.

Theorem (Black-box Uniform Stability):
loss $\ell : \mathbb{R}^d \to \mathbb{R}$ convex and L-smooth w.r.t. $\|\cdot\|_p$ and an optimization alg. $\mathcal{A}$ with convergence rate $C\hat{L}\|x_0 - x^*\|_p^{\hat{p}_1}/\hat{T}^\gamma$ for convex, $(\hat{L}, \hat{p}_1)$-Hölder smooth functions, where $\hat{p}_1 = \min\{p, 2\}$. Then, the iterate x_T produced by the black-box scheme with restarts $\mathrm{USOLP}(\mathcal{A}, T)$ initialized at x_0, satisfies,

1. $\varepsilon_{\text{stab}}(\mathrm{USOLP}(\mathcal{A}, T)) = \widetilde{\mathcal{O}}_p((T^\gamma/n)^{\frac{1}{\hat{p}_2 - 1}} LR^2)$,
2. $\varepsilon_{\text{opt}}(\mathrm{USOLP}(\mathcal{A}, T)) \coloneqq f_S(x_T) - f_S(x^*) = \widetilde{\mathcal{O}}_p(LR^2/T^\gamma)$,

for $p \in (1, \infty)$ where $\hat{p}_2 = \max\{p, 2\}$ and $T = \sum_{i=1}^r \hat{T}_i$ is the sum of gradient oracle calls from all stages.

	LB (Non-Euclidean, ℓ_p-norm)	UB (Euclidean ℓ_2-norm)	UB (Non-Euclidean, ℓ_p-norm)
$d \leq n$	$\widetilde{\Omega}_p(LR^2\frac{d^{1/2-1/\hat{p}}}{n^{1/2}})$ (Levy & Duchi; 2019)	$\widetilde{\mathcal{O}}_p(LR^2\left(\frac{1}{n}\right)^{1/2})$	$\widetilde{\mathcal{O}}_p(LR^2\frac{d^{1/2-1/\hat{p}}}{n^{1/2}})$ (This work)
$d > n$	$\widetilde{\Omega}_p(LR^2\left(\frac{1}{n}\right)^{1/\hat{p}})$ (This work)	(Attia & Koren; 2022)	$\widetilde{\mathcal{O}}_p(LR^2\left(\frac{1}{n}\right)^{1/\hat{p}})$ (This work)

Table 1. Excess risk bounds for black-box reduction algorithms for ERM with loss functions in $\mathbb{R}^d$ that are L-smooth over the ball of radius R w.r.t. ℓ_p-norm, $p \geq 1$, $\hat{p} = \max\{p, 2\}$.

- **Lower bound:** we extend low-dim ($d \leq n$) results in (Levy and Duchi; 2019) and improve upon high-dim ($d \geq n$) results in (Sridharan; 2012).
- **Upper bound:** we prove optimal stability using dimension-independent constants for uniform convexity (also giving speed-up captured by restarts)

Example: Classification in ℓ_p-balls

Example:

- **Data:** $(Z_i, Y_i) \in \mathbb{R}^d \times \{-1, 1\}$, with $\|Z_i\|_q \leq R$, $q \in (1, 2]$
- **Loss function:** $\ell(x, (z, y)) = h(y\langle x, z\rangle)$, $h : \mathbb{R} \to \mathbb{R}$ convex and L-smooth

The $\ell(\cdot, (z, y))$ loss is LD^2-smooth with respect to both

- the ℓ_p norm, with $p = \frac{q}{q-1} \geq 2$
- the ℓ_2 norm (as $\|x\|_2 \leq \|x\|_q$)

Which regularizer to use? Depends!

- Sample size dependence: ℓ_2 norm yields rates that are $n^{1/2-1/p}$ better
- Dimensionality dependence: ℓ_p norm yields rates that are up to $d^{1/2-1/p}$ better (Depending on the ℓ_p distance from initial point of algorithm to ERM minimizer).