گزارش پروژه

دانشکده مهندسی کامپیوتر

نام و نام خانوادگی دانشجو: سید احسان حسن بیگی

استاد: دکتر نجفی

پاییز ۱۴۰۳

فهرست مطالب

۱. مقدمه

در این گزارش به بررسی مقاله‌ی

Certifying Some Distributional Robustness with Principled Adversarial Training [1]

می‌پردازیم.

بسیاری از مسائلی که در یادگیری ماشین کلاسیک بررسی شده‌اند، بر این فرض استوار اند که توزیع داده های آموزشی و داده های تست یکسان است. حال آن که در اکثر شرایط دنیای واقعی این فرض برقرار نمی‌باشد. به طور کلی مقاوم بودن مدل ها نسبت به تغییر توزیع (distribution shift) ، یک ویژگی مطلوب به حساب می‌آید و بنابراین در حوزه های مختلف به انواع distribution shift پرداخته شده است. حملات خصمانه یکی از شاخص ترین انواع distribution shift به حساب می‌آیند. روش کار این حملات به این صورت است که فرد متخاصم با به دست آوردن یک نویز کوچک طراحی شده و اضافه کردن آن به نمونه های ورودی، مدل را فریب می‌دهد تا نتیجه‌ی اشتباهی را خروجی دهد. کار های اولیه‌ای که به این موضوع پرداخته بودند نشان دادند که اکثر مدل ها نسبت به این حملات آسیب پذیر اند و به دست آوردن نویز ذکر شده نیز کار ساده‌ای است. به صورت شهودی می‌توان گفت که مرز تصمیم اکثر مدل ها، به دلیل قرار داشتن در ابعاد بالا، به تمامی داده‌ها نزدیک است و بنابراین حتی با ایجاد فاصله‌ی کم نسبت به نمونه های آموزش نیز می‌توان از مرز تصمیم عبور کرد. به طور کلی پیدا کردن جهت تغییرات نسبت به اندازه‌ی تغییرات از اهمیت بیشتری برخوردار است و جهت تغییرات نیز به سادگی، توسط گرفتن مشتق تابع هزینه نسبت به ورودی مدل و اعمال gradient ascent بر روی آن قابل محاسبه است. ضمنا با توجه به اینکه خروجی اکثر مدل ها نسبت به ورودی‌شان piecewise linear است (به دلیل استفاده از ReLU)، می‌توان تنها با تعداد کمی iteration از gradient ascent به نتیجه‌ی مطلوب رسید.

با توجه به گستردگی این نوع آسیب پذیری، روش های متنوعی هم برای حمله و هم برای دفاع ارائه شده است. به منظور مقاوم سازی مدل نسبت به حملات خصمانه روندی برای آموزش به نام adversarial training ارائه شده است که به صورت زیر تعریف می‌شود:

$$\inf_{\theta \in \Theta} \sup_{q \in B_\varepsilon(p)} \mathbb{E}_q[l(\theta; z)]$$

به این معنا که آپدیت کردن پارامتر های شبکه در هر مرحله، پس از انتخاب بدترین توزیع ممکن توسط فرد متخاصم انجام شود. بدیهی است که استفاده از این روند برای آموزش مدل، به مقاوم سازی آن در برابر حملات کمک می‌کند.

نکات زیر در رابطه با عبارت ذکر شده حائز اهمیت است:

- بودجه‌ی فرد متخاصم برای انتخاب بدترین توزیع محدود است. به این صورت که توزیع q انتخاب شده توسط فرد متخاصم، باید در ε همسایگی توزیع اصلی p باشد.
- عبارت ذکر شده فرم کلی تر adversarial training است که به فرد متخاصم distributional adversary می‌گویند. در این حالت فرد متخاصم می‌تواند تمام بودجه اش را به نحوه‌ی دلخواه خودش روی توزیع اعمال کند. حالت محدود تری نیز برای فرد متخاصم وجود دارد که به آن pointwise adversary گفته می‌شود و به صورت زیر است:

$$\inf_{\theta\in\Theta} \mathbb{E}_{z\sim p}\left[\sup_{u:\ \|u\|_p\le\varepsilon} l(\theta; z+u)\right]$$

در این حالت بودجه‌ی تغییرات به ازای هر نمونه تعریف می‌شود. در این مقاله بر روی حالت کلی تر distributional robustness صحبت می‌شود.

- عبارت ذکر شده معادل یک بازی minimax است که محاسبه‌ی آن NP-hard است. بنابراین روش‌های موجود به جای محاسبه‌ی supremum درونی، از heuristic هایی استفاده می‌کنند که آن را تقریب بزنند.

در این مقاله روش جدیدی برای دفاع در برابر حملات خصمانه ارائه شده است که ادعا می‌کند در شرایطی که مقاوم سازی زیادی مورد نیاز نباشد (ε همسایگی کوچک باشد) و تابع هزینه نیز هموار باشد، می‌توان فرم معادلی برای adversarial training ارائه داد که به صورت بهینه قابل محاسبه است (computationally tractable) و از نظر پیچیدگی محاسباتی مرتبه‌ی زمانی آن مشابه روند عادی آموزش (ERM) است. همچنین یک کران بالا نیز برای بدترین حالت تابع هزینه ارائه می‌دهند. به این معنا که در صورت آموزش طبق روند ارائه شده، در بدترین حالت نیز عملکرد مدل از این کران بهتر خواهد بود.

۲. روش ارائه شده

۲.۱. روند distributionally robust optimization (DRO)

همان طور که گفته شد، این مقاله به فرم کلی adversarial training می‌پردازد که مسئله‌ای از فرم DRO است. فرم اصلی DRO به صورت زیر است:

$$\inf_{\theta\in\Theta} \sup_{q\in B_\varepsilon(p)} \mathbb{E}_q[l(\theta;z)]$$

در کار های پیشین یک فرم dual برای supremum داخلی ارائه شده که به صورت زیر می‌باشد:

$$\sup_{q\in B_\varepsilon(p)} \mathbb{E}_q[l(\theta;z)] = \inf_{\gamma\geq 0} \gamma\varepsilon + \mathbb{E}_{z\sim p}\left[\sup_{z'} l(\theta;z') - \gamma c(z,z')\right]$$

حال اگر به جای توزیع p توزیع تجربی $\hat{\mathrm{p}}_n$ را جایگذاری کنیم خواهیم داشت:

$$\inf_{\theta\in\Theta} \inf_{\gamma\geq 0} \gamma\varepsilon + \mathbb{E}_{z\sim\hat{p}_n}\left[\sup_{z'} l(\theta;z') - \gamma c(z,z')\right]$$

حال با توجه به اینکه در عمل ترجیح می‌دهیم γ را توسط cross validation به دست بیاوریم، می‌توان آن را از عبارت بالا حذف کرد و به فرم lagrangian relaxation از DRO رسید:

$$F(\theta) \coloneqq \inf_{\theta\in\Theta} \mathbb{E}_{Z\sim\hat{p}_n}[\varphi_\gamma(\theta;\, Z)]$$

$$\varphi_\gamma(\theta;\, z) \coloneqq \sup_{z'} l(\theta;\, z') - \gamma c(z', z)$$

حال اگر دقت کنیم، در عبارتی که باید بهینه کنیم، هم لاس به θ وابسته است و هم داده‌های خصمانه (z') یعنی:

$$F(\theta) = \inf_{\theta\in\Theta} \mathbb{E}_{z\sim\hat{p}_n}\left[l\left(\theta; z'^*(\theta)\right) - \gamma c\left(z, z'^*(\theta)\right)\right]$$

بنابراین گرفتن مشتق این عبارت نسبت به θ کار ساده ای نیست و باید از envelope theorem استفاده کرد. اما طبق این قضیه ابتدا باید به ازای یک z ثابت

$$\mathrm{T}_\gamma(\theta;\, z) \coloneqq \operatorname*{argmax}_{z'} l(\theta;z') - \gamma c(z,z')$$

را محاسبه کنیم که این کار نیز قابل انجام نیست زیرا این عبارت لزوما concave نمی‌باشد.

در این مرحله، برای ادامه‌ی محاسبات نیازمند اعمال قید هایی هستیم که به واسطه‌ی آن بتوان عبارت ذکر شده را محاسبه کرد. به طور کلی نحوه‌ی انتخاب مجموعه‌ی $Q = \{q: q \in B_\varepsilon(p)\}$، تعیین کننده‌ی میزان مقاومت به دست آمده در برابر حملات و همچنین قابل محاسبه بودن مسئله‌ی بهینه سازی می‌باشد. ایده‌ی اصلی این مقاله آن است که با چند فرض ساده، روندی برای آموزش ارائه می‌دهد که از نظر بار محاسباتی بهینه است و ضمنا نسبت به کار های پیشین، قید های کمتری نیاز دارد.

- **قید ۱**: تابع هزینه‌ی c باید پیوسته بوده و به ازای هر z_0 تابع $c(\cdot, z_0)$ باید نسبت به نورم 1-strongly convex باشد
- **قید ۲**: تابع $l(\theta;\ z)$ باید قید های Lipschitzian smoothness را داشته باشد:

$$\|\nabla_\theta l(\theta;\ z) - \nabla_\theta l(\theta';\ z)\|_* \leq L_{\theta\theta}\ \|\theta\ -\ \theta'\|$$

$$\|\nabla_\theta l(\theta;\ z) - \nabla_\theta l(\theta;\ z')\|_* \leq L_{\theta z}\ \|z\ -\ z'\|$$

$$\|\nabla_z l(\theta;\ z) - \nabla_z l(\theta';\ z)\|_* \leq L_{z\theta}\ \|\theta\ -\ \theta'\|$$

$$\|\nabla_z l(\theta;\ z) - \nabla_z l(\theta;\ z')\|_* \leq L_{zz}\ \|z\ -\ z'\|$$

در عمل برای برقراری این قید باید از activation function های هموار مانند sigmoid, ELU استفاده کرد و مواردی مانند ReLU این شرط را ارضا نمی‌کنند.

می‌توان نشان داد که با فرض وجود قید های ۱ و ۲، اگر γ به اندازه‌ی کافی بزرگ باشد (یعنی $\gamma \geq L_{zz}$ که به معنای ε کوچک است) آنگاه φ_γ نیز هموار خواهد بود و عبارت $l(\theta; z') - \gamma c(z, z')$ نسبت به z' به صورت strictly concave خواهد شد و بنابراین عبارت ذکر شده قابل محاسبه می‌شود. البته محاسبه‌ی این مقدار نیز به صورت دقیق مقدور نبودن و با الگوریتم gradient ascent آن را تقریب می‌زنیم.

به طور خلاصه می‌توان گفت که در روند پیشنهادی برای آموزش، هدف ما محاسبه‌ی $F(\theta)$ است که با فرض برقراری قید های ۱ و ۲ (هموار بودن تابع لاس نسبت به z) و همچنین به ازای γ بزرگ (مقاوم سازی کم تا متوسط) می‌توان این مقدار را به صورت بهینه محاسبه کرد. طبق envelope theorem ابتدا توسط چند گام gradient ascent مقدار $T_\gamma(\theta; z)$ را به دست می‌آوریم و سپس توسط یک گام gradient descent در جهت کمینه کردن

$$\mathbb{E}_{Z\sim\hat{p}_n}\left[\varphi_\gamma(\theta;\ Z)\right]$$

گام برمی‌داریم و θ را به روز رسانی می‌کنیم. تکرار این روند همان فرایند آموزش پیشنهادی است.

شایان ذکر است که به واسطه‌ی استفاده از فرم lagrangian relaxation، می‌توان نشان داد که gradient ascent درونی به صورت خطی همگرا می‌شود.

۲.۲. ارائه‌ی certificate of robustness

این مقاله یک کران بالای وابسته به داده نیز برای بدترین حالت تابع لاس ارائه می‌دهد. به این معنا که در بدترین حالت نیز عملکرد مدل از این کران بالا بهتر خواهد بود. طبق فرم dual ذکر شده داریم:

$$\sup_{q\in B_\varepsilon(p)} \mathbb{E}_q[l(\theta;\ z)] = \inf_{\gamma\geq 0}\ \gamma\varepsilon + \mathbb{E}_{Z\sim\hat{p}_n}\left[\varphi_\gamma(\theta;\ Z)\right]$$

$$\rightarrow \sup_{q\in B_\varepsilon(p)} \mathbb{E}_q[l(\theta;\ z)] \leq \gamma\varepsilon + \mathbb{E}_{Z\sim\hat{p}_n}\left[\varphi_\gamma(\theta;\ Z)\right]$$

حال طبق قضیه‌ی ۳ مقاله اگر $|l(\theta;\ z)| \le M_l$ آنگاه به ازای یک t ثابت و $b_1, b_2 > 0$ به احتمال حداقل $1 - e^{-t}$ به طور همزمان برای تمام $\theta, \varepsilon \ge 0, \gamma \ge 0$ داریم:

$$\sup_{q \in B_\varepsilon(p)} \mathbb{E}_q[l(\theta; z)] \le \gamma\varepsilon + \mathbb{E}_{Z\sim\hat{p}_n}\left[\varphi_\gamma(\theta;\ Z)\right] + \varepsilon'_{n(t)}$$

$$\varepsilon'_n(t) \coloneqq \gamma b_1 \sqrt{\frac{M_l}{n}} \int_0^1 \sqrt{\log\ N(F, M_l\,\varepsilon', \|\cdot\|_{L^\infty(Z)})}\ d\varepsilon' + b_2 M_l \sqrt{\frac{t}{n}}$$

$$\mathrm{F} \coloneqq \{\mathrm{l}(\theta,\cdot) : \theta \in \Theta\}$$

و $N(V, \varepsilon', \|\cdot\|)$ نیز covering number مجموعه V نسبت به نورم $\|\cdot\|$ است. به طور خاص اگر $\varepsilon = \hat{\varepsilon}_n(\theta)$ باشد، این کران به صورت

$$\sup_{q \in B_{\hat{\varepsilon}_n(\theta)}(p)} \mathbb{E}_q[l(\theta; z)] \le \gamma\hat{\varepsilon}_n(\theta) + \mathbb{E}_{Z\sim\hat{p}_n}\left[\varphi_\gamma(\theta;\ Z)\right] + \varepsilon'_{n(t)}$$

$$= \sup_{q \in B_{\hat{\varepsilon}_n(\theta)}(p)} \mathbb{E}_q[l(\theta;\ z)] + \varepsilon'_{n(t)}$$

در می‌آید که tight است.

بنابراین مطابق قضیه‌ی ۳ مقاله، در حالت کلی $\varepsilon'_{n(t)}$ به صورت خطی با d اسکیل می‌شود. حال اگر قید مربوط به Lipschitz بودن تابع لاس را نیز اعمال کنیم، مطابق corollary 1 مقاله داریم:

$$\varepsilon'_n(t) = b_1 \sqrt{\frac{d\left(L\ diam(\Theta) + M_{\theta_0}\right)}{n}} + b_2\left(L\ diam(\Theta) + M_{\theta_0}\right)\sqrt{\frac{t}{n}}$$

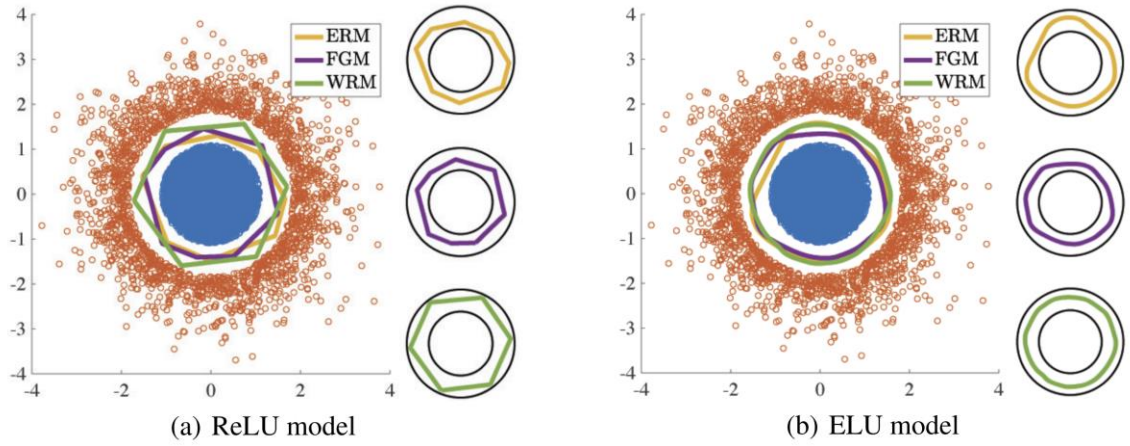
که نسبت به قبل کران بهتری است.

همچنین نشان داده می‌شود که adversarial perturbation های استفاده شده در روند آموزش، قابلیت تعمیم پذیری دارند. به این معنا که روند آموزش ارائه شده با اینکه از توزیع تجربی داده ها استفاده می‌کند، اما مقاوم سازی ای که نتیجه می‌دهد، معادل استفاده از توزیع واقعی داده هاست.

۳. آزمایش عملکرد

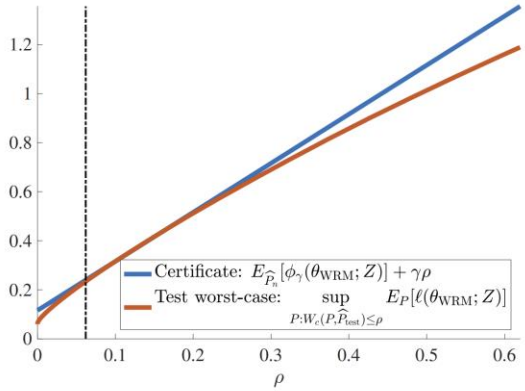
به منظور آزمایش روش ارائه شده، یک دیتاست synthetic با فضای ویژگی دو بُعدی ایجاد شده است که شامل دو کلاس می‌باشد. نمونه‌ی های مربوط به کلاس اول در دایره‌ای به شعاع $\sqrt{2}/1.3$ قرار دارند و نمونه های مربوط به کلاس دوم نیز خارج دایره ای به شعاع $\sqrt{2} \times 1.3$ قرار دارند تا دسته بندی جدایی پذیر باشد. در ادامه به توضیح نتایج آزمایش های انجام شده می‌پردازیم:

- در تصویر زیر روش ارائه شده (WRM) نسبت به دو روش دیگر (FGM, ERM) سنجیده می‌شود. همچنین دو تابع فعال سازی ReLU, ELU نیز برای هر کدام از سه روش با یکدیگر مقایسه می‌شوند. می‌دانیم که به طور کلی هر چه مرز تصمیم متقارن تر و در وسط دو کلاس باشد، مدل به دست آمده مقاوم تر است زیرا هر بخش از مرز تصمیم که به یک دسته نزدیک تر باشد، مستعد exploit شدن توسط حملات خصمانه می‌باشد. کافی است نویزی به ورودی اضافه شود که کمی آن را در جهت ذکر شده تغییر دهد و مدل به ازای این ورودی فریب خواهد خورد. همان طور که مشاهده می‌شود مرز تصمیم برای روش ارائه شده نسبت به دو روش دیگر متقارن تر است. همچنین مشاهده می‌شود که استفاده از تابع فعال سازی هموار مانند ELU، مقاومت بهتری نسبت به ReLU به دست می‌دهد. دلیل این مورد نیز قید های هموار بودن تابع لاس است که پیش تر ذکر شد.

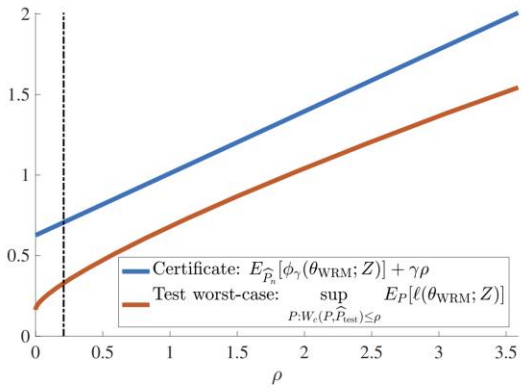


(a) ReLU model (b) ELU model

- تصویر زیر به بررسی certificate of robustness می‌پردازد. همان طور که مشاهده می‌شود به ازای ε های مختلف خطا همواره از کران بالای ارائه شده کمتر است. همچنین خط عمودی نقطه ای را نشان می‌دهد که $\varepsilon = \hat{\varepsilon}_n(\theta)$ و مشاهده می‌شود که مطابق ادعای انجام شده، در این نقطه، کران بالای آن tight خواهد بود.

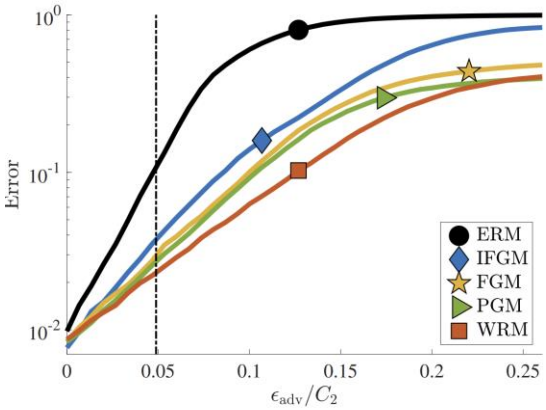


(a) Synthetic data

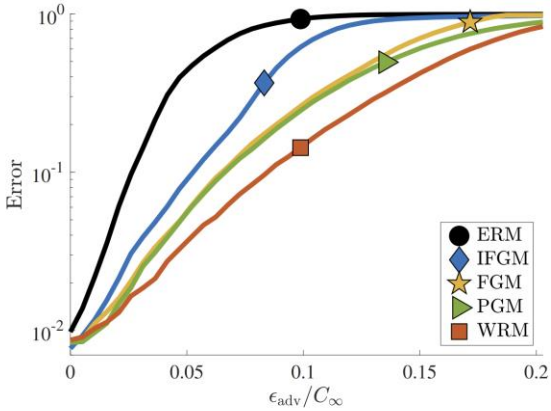


(b) MNIST

- تصویر زیر خطای misclassification روش های مختلف را نسبت به میزان بودجه‌ی perturbation حمله (حملات نورم دو و نورم بی‌نهایت) رسم کرده است و مشاهده می‌شود که خطای روش WRM نسبت به روش های دیگر کمتر بوده و مدل مقاوم تری به دست آمده است.

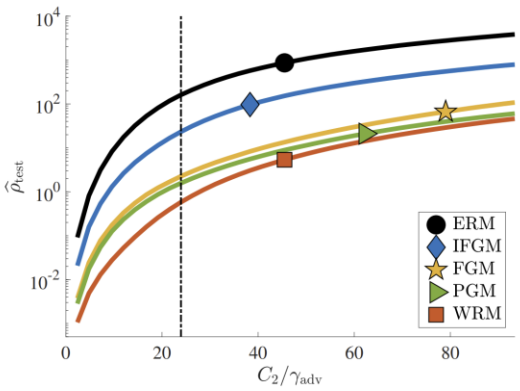


(a) Test error vs. ϵ_{adv} for $\|\cdot\|_2$ attack

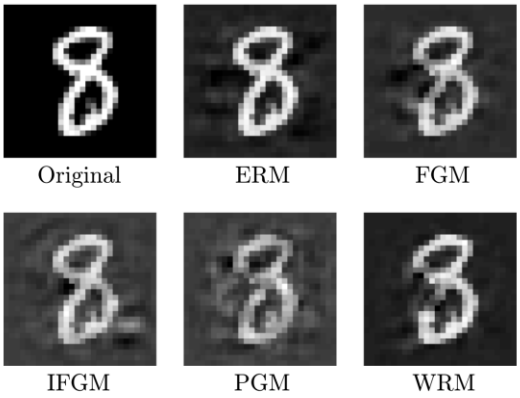


(b) Test error vs. ϵ_{adv} for $\|\cdot\|_\infty$ attack

- در تصویر زیر، بخش اول، فاصله‌ی توزیع perturbed به دست آمده نسبت به توزیع اصلی به ازای هر γ رسم شده است که می‌تواند ثبات محلی مرز تصمیم برای ورودی ها را نشان دهد و خط عمودی نیز γ استفاده شده در زمان آموزش را نشان می‌دهد. با توجه به اینکه محور عمودی نمودار به صورت log scale است، پایین تر بودن خم مربوط به روش WRM نسبت به سایر روش ها نشان دهنده‌ی ثبات بیشتر آن است. همچنین در بخش دوم، کمترین میزان perturbation مورد نیاز برای فریب دادن مدل به ازای هر روش رسم شده است و مشاهده می‌شود که برای روش WRM به تغییرات بیشتری نیاز است و بنابراین این روش مقاومت بیشتری ایجاد می‌کند.



(a) $\widehat{\rho}_{test}$ vs. $1/\gamma_{\text{adv}}$



(b) Perturbations on a test datapoint

۴. نتیجه گیری

همان طور که عنوان شد، فرم کلی مسئله‌ی DRO با این که باعث مقاوم سازی مدل در برابر حملات خصمانه می‌شود اما به دلیل قابل محاسبه نبودن supremum داخلی، قابل دست یابی نیست. به این منظور کار های پیشین عمدتا به جای محاسبه‌ی بدترین توزیع خصمانه توسط supremum، از heuristic هایی استفاده می‌کردند. کار های دیگری نیز بودند که به صورت دقیق روش شان را اثبات می‌کردند اما برای این منظور فرض های محدود کننده ای اعمال می‌شد. در این کار روندی برای انجام adversarial training ارائه شده است که هم دارای گارانتی های دقیق آماری است، هم از نظر بار محاسباتی بهینه است و نرخ همگرایی مناسبی دارد و هم برای خانواده‌ی بزرگ تری از مسائل قابل اعمال می‌باشد.

قید های مورد نیاز برای برقراری نتایج این مقاله همچنان خانواده‌ی بزرگی از مسائل را شامل می‌شوند و از نگاه کاربردی، تنها محدودیت هایی برای استفاده از توابع فعال سازی و میزان مقاوم سازی مدل وجود دارد. تئوری های ارائه شده و همچنین نتایج آزمایش های انجام شده، نشان می‌دهد که استفاده از شبکه های هموار (به عنوان مثال استفاده از تابع فعال سازی ELU به جای ReLU) از نظر مقاوم سازی مدل دارای مزیت است. شایان ذکر است که در صورت برقرار نبودن این قید ها، همچنان روش ارائه شده موثر خواهد بود اما گارانتی های ذکر شده دیگر برقرار نمی‌باشند و این روند به یک heuristic تقلیل می‌یابد.

۵. مراجع

[1] Sinha, A., Namkoong, H., & Duchi, J. (2018, February). Certifying Some Distributional Robustness with Principled Adversarial Training. In *International Conference on Learning Representations*.