Lipschitz continuity and neural networks

Dávid Terjék
dterjek@renyi.hu

Alfréd Rényi Institute of Mathematics
Budapest, Hungary

MILAB Deep Learning seminar, 2020

Lipschitz continuity

$f : (X, d_X) \to (Y, d_Y)$ is λ-Lipschitz if

$$d_Y(f(x_1), f(x_2)) \leq \lambda \cdot d_X(x_1, x_2)$$

holds for $\forall x_1, x_2 \in X$. The smallest such λ is the Lipschitz norm

$$\|f\|_L = \sup_{x_1, x_2 \in X; x_1 \neq x_2} \frac{d_Y(f(x_1), f(x_2))}{d_X(x_1, x_2)},$$

which quantifies how much f can dilate distances.

Outline

Why $\|f\|_L$ matters when f is a neural network, and how to estimate $\|f\|_L$ or enforce $\|f\|_L \leq \lambda$?

Why?

- Adversarial methods
  - Wasserstein GAN
  - Mutual Information Neural Estimation
- Generalization, robustness, stability

How?

- Lipschitz regularization
  - Penalty methods
  - Normalization methods
- Lipschitz constant estimation

Why?

Wasserstein metric

$$\mu, \nu \in P(X):$$

$$\pi \in \Pi(\mu, \nu) \subset P(X \times X)$$

$$\Longleftrightarrow$$

$$\forall A \subset X : \pi(A \times X) = \mu(A) \wedge \pi(X \times A) = \nu(A)$$

$$W(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \mathbb{E}_{(x_1, x_2) \sim \pi} d(x_1, x_2)$$

$\Longrightarrow$ $(P(X), W)$ is a metric space!

Wasserstein GAN

Kantorovich-Rubinstein:

$$W(\mu,\nu) = \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(x_1,x_2)\sim\pi} d(x_1,x_2) = \sup_{\|f\|_L \le 1} \mathbb{E}_{x\sim\mu} f(x) - \mathbb{E}_{x\sim\nu} f(x)$$

Wasserstein GAN [Arjovsky et al., 2017]:

$$\min_{\theta_g \in \mathbb{R}^m} \max_{\theta_f \in \mathbb{R}^n, \|f(\cdot,\theta_f)\|_L \le 1} \mathbb{E}_{x\sim\mu} f(x,\theta_f) - \mathbb{E}_{z\sim\zeta} f(g(z,\theta_g),\theta_f)$$

$\implies$ *gradient vanishing* solved

Non-Wasserstein GANs with $\|f(\cdot,\theta_f)\|_L \le \lambda$ [Zhou et al., 2019]:
$\implies$ *gradient uninformativeness* solved

How to enforce $\|f(\cdot,\theta_f)\|_L \le \lambda$*?*

Unconstrained Wasserstein GAN

$\alpha > 1$:

$$W(\mu,\nu)^\alpha = \sup_{f \in Lip(X)} \mathbb{E}_{x\sim\mu} f(x) - \mathbb{E}_{x\sim\nu} f(x) - (\alpha-1)\alpha^{\frac{\alpha}{1-\alpha}} \|f\|_L^{\frac{\alpha}{\alpha-1}}$$

Unconstrained Wasserstein GAN:

$$\min_{\theta_g \in \mathbb{R}^m} \max_{\theta_f \in \mathbb{R}^n} \mathbb{E}_{x\sim\mu} f(x,\theta_f) - \mathbb{E}_{z\sim\zeta} f(g(z,\theta_g),\theta_f) - (\alpha-1)\alpha^{\frac{\alpha}{1-\alpha}} \|f(\cdot,\theta_f)\|_L^{\frac{\alpha}{\alpha-1}}$$

How to estimate $\nabla_{\theta_f} \|f(\cdot,\theta_f)\|_L$*?*

Mutual Information

Radon-Nykodim:

$$\mu, \nu \in P(X)$$

$$\forall A \subset X : \nu(A) = 0 \implies \mu(A) = 0$$

$$\implies$$

$$\exists \frac{d\mu}{d\nu} : X \to \mathbb{R} \text{ s.t. } \forall A \subset X : \mu(A) = \int_A \frac{d\mu}{d\nu} d\nu$$

Kullback-Leibler:

$$D(\mu \| \nu) = \mathbb{E}_{x \sim \mu} \frac{d\mu}{d\nu}(x)$$

Mutual Information:

$$I(X, Y) = D(\mu_{XY} \| \mu_X \times \mu_Y)$$

Mutual Information Neural Estimation

Donsker-Varadhan:

$$D(\mu\|\nu) = \mathbb{E}_{x\sim\mu}\frac{d\mu}{d\nu}(x) = \sup_{f\in B(X)} \mathbb{E}_{x\sim\mu} f(x) - \log \mathbb{E}_{x\sim\nu} e^{f(x)}$$

Mutual information maximization [Belghazi et al., 2018]:

$$\min_{\theta_g\in\mathbb{R}^m} \max_{\theta_f\in\mathbb{R}^n} \mathbb{E}_{x\sim\mu} f((x, g(x,\theta_g)), \theta_f) - \log \mathbb{E}_{x_1\sim\mu, x_2\sim\mu} e^{f((x_1, g(x_2,\theta_g)),\theta_f)}$$

Additional Lipschitz constraint $\|f(\cdot,\theta_f)\|_L \leq \lambda$ [Ozair et al., 2019]:
$\implies$ *high sample complexity* solved

Moreau-Yosida-Kullback-Leibler divergence

Moreau-Yosida:

$$F_\lambda(x) = \inf_y F(y) + \lambda d(x,y)$$

$$F : (X,d) \to [0,\infty) \text{ lsc } \implies F_\lambda(x) \xrightarrow{\lambda\to\infty} F(x) \wedge \|F_\lambda\|_L \leq \lambda$$

Applied to $(\mu \to D(\mu\|\nu)) : (P(X), W) \to [0,\infty)$:

$$D_\lambda(\mu\|\nu) := \inf_{\xi\in P(X)} D(\xi\|\nu) + \lambda W(\mu,\xi)$$

$$= \sup_{\|f\|_L\leq\lambda} \mathbb{E}_{x\sim\mu} f(x) - \log \mathbb{E}_{x\sim\nu} e^{f(x)}$$

$$D_\lambda(\mu\|\nu) \xrightarrow{\lambda\to\infty} D(\mu\|\nu) \wedge \|(\mu \to D_\lambda(\mu\|\nu))\|_L \leq \lambda$$

Generalization, robustness, stability

Generalization theory of deep neural networks [Bartlett et al., 2017, Wei and Ma, 2020]: Lipschitz continuity is a key component for proving generalization error bounds.

Adversarial robustness [Tsuzuku et al., 2018]: robustness certificates can be given based on Lipschitz continuity properties.

Model-based reinforcement learning [Asadi et al., 2018]: a Lipschitz continuous transition function implies a Lipschitz continuous estimated value function and error bounds for both value estimation and multi-step prediction.

How to estimate or upper bound $\|f(\cdot, \theta_f)\|_L$*?*

How?

Lipschitz regularization

Lipschitz regularization of neural networks divides into two main approaches.

One is to quantify the violation of the Lipschitz condition to be enforced by a data-dependent *penalty*, which is then added to the training objective.

The other includes *normalization* techniques for weight matrices and Lipschitz continuous activation functions, mostly based on the composition property $\|f_2 \circ f_1\|_L \leq \|f_1\|_L \cdot \|f_2\|_L$.

Gradient penalty

Rademacher:

$$\|f\|_L < \infty \implies \|(x \to \|\nabla_x f(x)\|_2)\|_\infty = \|f\|_L$$

Gradient penalty:

$$\max\{\|\nabla_x f(x)\|_2 - \lambda, 0\}$$

Wasserstein GAN with gradient penalty [Gulrajani et al., 2017]:

$$\min_{\theta_g \in \mathbb{R}^m} \max_{\theta_f \in \mathbb{R}^n} \mathbb{E}_{x\sim\mu} f(x, \theta_f) - \mathbb{E}_{z\sim\zeta} f(g(z, \theta_g), \theta_f)$$
$$- \ell \cdot \mathbb{E}_{x\sim\rho} \left(\max\{\|\nabla_x f(x, \theta_f)\|_2 - 1, 0\}\right)^2$$

Lipschitz penalty

Lipschitz penalty:

$$\max\left\{\frac{|f(x_1)-f(x_2)|}{\|x_1-x_2\|_2}-\lambda,0\right\}$$

Wasserstein GAN with Lipschitz penalty [Petzka et al., 2018]:

$$\min_{\theta_g\in\mathbb{R}^m}\max_{\theta_f\in\mathbb{R}^n}\mathbb{E}_{x\sim\mu}f(x,\theta_f)-\mathbb{E}_{z\sim\zeta}f(g(z,\theta_g),\theta_f)$$
$$-\ell\cdot\mathbb{E}_{(x_1,x_2)\sim\rho}\left(\max\left\{\frac{|f(x_1,\theta_f)-f(x_2,\theta_f)|}{\|x_1-x_2\|_2}-1,0\right\}\right)^2$$

$\implies$ divergent training

Adversarial Lipschitz penalty

$$\|f\|_L = \sup_{d(x,x+r)>0} \left\{ \frac{|f(x) - f(x+r)|}{d(x, x+r)} \right\}$$

Lipschitz adversarial perturbation:

$$r_{adv}(x) = \arg\max_{d(x,x+r)>0} \left\{ \frac{|f(x) - f(x+r)|}{d(x, x+r)} \right\}$$

Adversarial Lipschitz penalty:

$$\max \left\{ \frac{|f(x) - f(x + r_{adv}(x))|}{\|r_{adv}(x)\|_2} - \lambda, 0 \right\}$$

Wasserstein GAN with adversarial Lipschitz penalty [Terjék, 2020]:

$$\rho = (x \to (x, x + r_{adv}(x)))_{\#} \frac{1}{2}(\mu + g(\cdot, \theta_g)_{\#}\zeta)$$

$\implies$ convergent training

Approximation of $r_{adv}(x)$

$$\forall x \in \mathbb{R}^n : r \to |f(x) - f(x+r)| : \mathbb{R}^n \to \mathbb{R}$$

has a global minimum at $r = 0$, implying that

$$\nabla_r |f(x) - f(x+r)|(0) = 0,$$

so the 2nd order Taylor approximation at $r = 0$ is

$$|f(x) - f(x+r)| \approx \frac{1}{2} r \cdot \text{Hess}_r\, |f(x) - f(x+r)|(0) \cdot r^T,$$

which is locally maximized by the first eigenvector of the Hessian. Power iteration with Hessian-vector products converges to the direction of greatest change in $f(x)$ at x.

$r_{adv}(x)$ is then approximated by a random magnitude perturbation towards this adversarial direction.

Weight clipping

$$\theta_f \in K \subset \mathbb{R}^n, K \text{ compact} \implies \|f(\cdot, \theta_f)\|_L \leq \lambda(K)$$

Weight clipping [Arjovsky et al., 2017]:

$$\theta_f \in [-c, c]^n$$

Spectral normalization

$\|.\|_L$ of affine maps:

$$(x \to M \cdot x + b) : (\mathbb{R}^k, \|.\|_2) \to (\mathbb{R}^l, \|.\|_2)$$

$$\implies \|(x \to Mx + b)\|_L = \sigma_1(M)$$

Spectral normalization [Miyato et al., 2018]:

$$\overline{M} = \frac{M}{\sigma_1(M)}$$

Approximation of $\sigma_1(M)$:

$$v_{i+1} = \frac{1}{\|M^T u_i\|_2} M^T u_i, \; u_{i+1} = \frac{1}{\|M v_{i+1}\|_2} M v_{i+1}$$

$$\sigma_1(M) \approx u^T M v$$

Gradient norm preservation

Gradient norm attenuation:
$\|f\|_L = 1 \implies$ backpropagating a gradient through f can only decrease its norm, potentially resulting in nonlinear capacity being underused.

Gradient norm preserving architectures [Anil et al., 2019]:
Orthogonal weight matrices with GroupSort activations $\implies$ universal approximation of 1-Lipschitz functions.

Spectrally normalized ReLU NNs with GNP property *are linear*.

Orthogonalization

Orthogonality does not sacrifice capacity [Anil et al., 2019]:

Spectrally normalized layers can be replaced by layers with $\sigma_1(M) = \cdots = \sigma_k(M) = 1$, resulting in an equivalent NN $\implies$ orthogonalization effectively reduces the hypothesis space.

Lipschitz constant estimation

Exact computation of $\|f(\cdot, \theta_f)\|_L$ is NP-hard [Virmaux and Scaman, 2018, Jordan and Dimakis, 2020]

POP for upper bounds [Gómez et al., 2020]

Hierarchies of SDPs for increasingly tight upper bounds [Fazlyab et al., 2019, Chen et al., 2020]

MIP for exact computation, upper bounds if stopped early [Jordan and Dimakis, 2020]

References I

Anil, C., Lucas, J., and Grosse, R. B. (2019).
Sorting out lipschitz function approximation.
In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 291–301. PMLR.

Arjovsky, M., Chintala, S., and Bottou, L. (2017).
Wasserstein generative adversarial networks.
In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR.

References II

Asadi, K., Misra, D., and Littman, M. L. (2018).
Lipschitz continuity in model-based reinforcement learning.
In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 264–273. PMLR.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. (2017).
Spectrally-normalized margin bounds for neural networks.
In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6240–6249.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Hjelm, R. D., and Courville, A. C. (2018).
Mutual information neural estimation.
In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 530–539. PMLR.

Chen, T., Lasserre, J., Magron, V., and Pauwels, E. (2020).
Polynomial optimization for bounding lipschitz constants of deep networks.
CoRR, abs/2002.03657.

References IV

Fazlyab, M., Robey, A., Hassani, H., Morari, M., and Pappas, G. J. (2019).
Efficient and accurate estimation of lipschitz constants for deep neural networks.
In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 11423–11434.

Gómez, F. L., Rolland, P., and Cevher, V. (2020).
Lipschitz constant estimation of neural networks via sparse polynomial optimization.
In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

References V

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017).
Improved training of wasserstein gans.
In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5767–5777.

Jordan, M. and Dimakis, A. G. (2020).
Exactly computing the local lipschitz constant of relu networks.
CoRR, abs/2003.01219.

References VI

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018).
Spectral normalization for generative adversarial networks.
In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Ozair, S., Lynch, C., Bengio, Y., van den Oord, A., Levine, S., and Sermanet, P. (2019).
Wasserstein dependency measure for representation learning.
In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 15578–15588.

References VII

Petzka, H., Fischer, A., and Lukovnikov, D. (2018).
On the regularization of wasserstein gans.
In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Terjék, D. (2020).
Adversarial lipschitz regularization.
In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tsuzuku, Y., Sato, I., and Sugiyama, M. (2018).
Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks.
In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31: Annual Conference on*

References VIII

Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada, pages 6542–6551.

Virmaux, A. and Scaman, K. (2018).
Lipschitz regularity of deep neural networks: analysis and efficient estimation.
In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 3839–3848.

Wei, C. and Ma, T. (2020).
Improved sample complexities for deep neural networks and robust classification via an all-layer margin.
In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

References IX

Zhou, Z., Liang, J., Song, Y., Yu, L., Wang, H., Zhang, W., Yu, Y., and Zhang, Z. (2019).
Lipschitz generative adversarial nets.
In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7584–7593. PMLR.