바서스테인 GAN 에서 칸트로비치 거리와 **KL**-발산에 관한 차이 분석

이우용, 김근영, 조동욱°
한국전자통신연구원 6G 무선방식연구실, 충북도립대학교°

{wylee, kykim12}@etri.re.kr, ducho@cpu.ac.kr

Difference analysis of Kantorovich-Wasserstein (KW) metric and Kullback-Leibler (KL)-divergence in Wasserstein GAN

Lee Woo Yong, Kim Keunyoung, and Cho Dongwook°

6G Wireless Technology Research Section
Telecommunication & Media Research Laboratory
Electronics and Telecommunications Research Institute (ETRI) and
Chungbuk Provincial University°

요 약

칸토로비치-바서스테인(KW: Kantorovich-Wasserstein) 거리(metric)는 바서스테인 GAN (Wasserstein Generative Adversarial Networks)과 같은 생성 모델을 학습하는 데 사용되는 손실(Loss) 함수이다. 기존의 바닐라(vanilla) GAN 손실에 비해 많은 장점이 있다. 이는 샘플 품질과 관련이 있고 최적화 과정의 향상된 안정성을 제공하며 Vanishing Gradient 문제를 겪지 않는다. 쿨백-레이블러(KL: Kullback-Leibler)-발산(divergence)과 칸토로비치-바서스테인(KW) 거리 관계가 규명되었다. KL-발산은 정보량을 엔트로피와 함께 정의되기 때문에, 정보 이론 분야에서 중요한 역할을 하고 있다. 그 동안 겉으로 보기에 관련이 없어 보이는 이 두 가지 KW 거리와 KL 발산에 어떤 공통점이 있는지 여부가 중요한 관심사였다. 최근 연구에서 최적 운송 문제(OTP: Optimal Transport Problem)가 암묵적 제약(조건)을 가짐으로써, 이 운송 문제는 전송률 왜곡과 정보 이론에서 연구된 최적 채널을 찾는 변분(variational) 문제와 관련될 수 있음이 입증되었다. 본 논문은 이들 결과를 바서스테인 GAN 에 손실함수로 적용했을 때, KL 발산과 칸토로비치 거리 관례에서 차이를 분석하고, 향후 씨멘틱(Semantic) 통신에서 정보량을 측정하기 위한 보다 효과적인 손실 함수 적용 방안을 검토한다.

Ⅰ. 서 론

칸토로비치-바서스테인(KW) 거리(metric, 길이)는 기계학습 생성 모델을 학습하는 데 널리 사용되는 손실(Loss) 함수이다[1, 2]. 기존의 바닐라(vanilla) GAN 손실[3]에 비해 많은 장점이 있다. 이는 샘플 품질과 관련이 있고 최적화 과정의 향상된 안정성을 제공하며 Vanishing Gradient 문제를 겪지 않는다[1]. 쿨백-레이블러(KL)-발산(divergence)과 칸토로비치-바서스테인(KW) 거리 관계가 규명되었다[4]. 가스파르 몽지(Gaspar Monge)에 의해 시작된 최적 운송 문제(OTP)에 대한 연구는 레오나드 칸토로비치(Leonid Kantorovich)가 확률 이론의 문제로 재구성하면서 크게 발전했다[4, 5]. 확률 변수(random variables) X와 Y를 두 개의 측도 가능한(measurable, 가측) 집합으로 설정하고, P(X)와 P(Y)를 각각 X와 Y에 대한 모든 확률 측도의 집합이라 하고, P(X×Y)를 X×Y에 대한 모든 결합 확률 측도 집합이라고 하자. c: X × Y → R을 음이 아닌(non-negative) 측도 가능 함수라고 할 때, 이를 비용 함수(cost function)라고 한다. 종종 X ≡ Y와 c(x, y)를 거리(metric)으로 사용한다. 확률 측도 ω ∈ P(X×Y)에 대한 기대 비용 $E_\omega\{c\}$은 다음 적분 값을 표현된다:

$$E_\omega\{c\} = \int_{X\times Y} c(x,y)\, d\omega(x,y)$$

여기서, 이 비용 함수를 위의 적분 값이 ω의 하반연속(lower semicontonuous) 또는 단힌 범함수(functional)라고 가정한다(즉, 집합 {ω: $E_\omega\{c\} \le \upsilon$}는 모든 υ ∈ R에 대해 닫힘). 특히, $c(\omega) = E_\omega\{c\}$가 연속 선형 범함수인 경우다.

Ⅱ. 본론: WGAN 에서 칸트로비치와 KL-발산 차이 분석

X 는 실제 영상이고 GAN 에 의해 생성된 페이크(Fake) 영상을 Y=G(Z)라고 할 때, 바서스테인 GAN 의 생성형 기계학습 구조는 그림 1 같다.

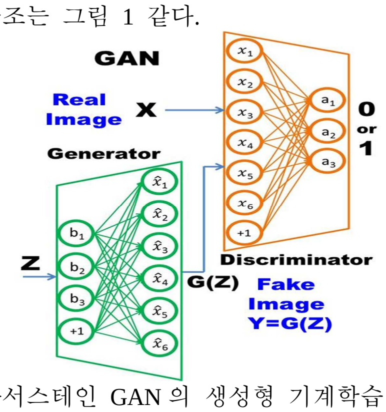

그림 1. 바서스테인 GAN 의 생성형 기계학습 구조.

비용 함수 c: X×Y → R 이 주어질 때, 모든 (x, y) ∈ X×Y 에 대하여 f(x) − g(y) ≤ c(x, y)의 조건을 만족하는 실수 함수 f: X → R 과 g: Y → R 를 고려하자. 그때 쌍대 공식 $J_c[p,q]$은 이 모든 함수에 대해 다음과 같은 최대화이다.

$$J_c[p,q] = sup\{E_p\{f\} - E_q\{g\}: f(x) - g(y) \le c(x,y)\}$$

여기서, 우리는 X≡Y 라고 가정한다. 다음과 같은 부등식을 만족함이 자명하다. 왜냐면 칸트로비치의 쌍대성을 만족하기 때문이다.

$$J_c[p,q] \le K_c[p,q]$$

KL-발산과 칸트로비치 거리 사이의 관계를 찾기 위해 이 쌍대 공식을 사용하려고 한다. 먼저, KL-발산의 다음과

같이 분해할 수 있다.

$$D[p,q] = D[p,r] + D(r,q) - \int_X \ln\frac{dq(x)}{dr(x)}[dq(x) - dr(x)]$$
$$= D[p,r] - D(q,r)$$
$$- \int_X \ln\frac{dq(x)}{dr(x)}[dp(x) - dq(x)]$$

이제 추가 제약(조건)을 만족하는 함수: f(x) − g(y) ≤ c(x, y)를 고려한다[5].

$$\beta f(x) = \nabla D[p,r] = \ln\frac{dp(x)}{dr(x)}, \beta \geq 0,$$

$$\alpha g(x) = \nabla D[q,r] = \ln\frac{dq(x)}{dr(x)}, \alpha \geq 0$$

정리 1(바서스테인 GAN 에서 칸트로비치 거리와 KL-발산 관계): 한 쌍의 함수 (f, g)을 쌍대 OTP 에 대한 해라고 하자. f = ▽D[p, r]와 g = ▽D[q, r]와 같은 기준 측도 r ∈ Q (Y)이 존재한다면, 임의의 $\epsilon \geq 0$ 에 대하여 바서스테인 GAN 에서 KL-발산 $I(X,Y)$와 이에 해당하는 칸트로비치 정보(거리) $K_c[\omega, q \otimes p]$는 다음 수식을 만족한다.

$$K_c[\omega, p \otimes q] \geq I(X,Y) + \epsilon.$$

요약 증명: 바서스테인 GAN 에서 x 는 실제 영상이고 GAN 에 의해 페이크(Fake) 영상을 y=G(z)을 생성한다고 하자. 바서스테인 GAN 의 정의 y=G(z)에 의하여 다음 식을 만족한다.

$$g(y) - g\big(G(z)\big) = 0$$

임의의 $\epsilon \geq 0$, $\epsilon' \geq 0$에 대하여, 바서스테인 GAN 은 생성된 영상 G(z)와 실제 영상 x 사이의 차이를 다음 식이 만족하도록 훈련된다.

$$sup|x - G(z)| \leq \epsilon$$
$$sup|f(x) - f\big(G(z)\big)| \leq \epsilon$$

그러므로, 실수함수 f: X → R 과 g: Y → R 에 대하여 다음 식을 만족한다.

$$sup|f(x) - f\big(G(z)\big) - g(y) + g\big(G(z)\big)| \leq \epsilon$$

생성된 영상 G(z)에 대하여, 비용 함수 c: X×Y → R 이 주어질 때 모든 (x, y) ∈ X×Y 에 대하여 f(x) − g(y) ≤ c(x, y)의 조건을 만족하고, 비용 함수 c 에 대하여 L_2-놈(norm)을 가정한다면 다음 식을 만족한다[2].

$$sup\big|f\big(G(z)\big) - g\big(G(z)\big)\big| = c\big(x, G(z)\big) = \|x - G(z)\|_2 \leq \epsilon$$

위 수식들로부터 다음 제한 조건을 만족한다.

$$sup|f(x) - g(y)| \leq \epsilon$$

따름 정리 2[6]의 수식의 우변 첫 번 째 $\kappa[f] - \kappa[g]$식은 다음 조건을 만족한다.

$$\kappa[f] - \kappa[g] = \ln\int_x \frac{dp(x)}{dr(x)}dr(x) - \ln\int_Y \frac{dq(y)}{dr(y)}dr(y)$$
$$= ln\frac{p(x)}{q(y)} = f(y) - g(y) \leq \epsilon$$

여기서, $\kappa(\cdot) = \ln\int_Y e^{(\cdot)}\,dr(y)$ 는 정규화 상수(누적 생성 함수의 값)이다. 따름 정리 2[6]의 수식의 우변 두 번 째 $\int_X g(x)\,[dp(x) - dq(x)]$식은 다음 조건을 만족한다.

$$dp(y) - dq(y) = \big(e^{f(y)-\kappa[f]} - e^{g(y)-\kappa[g]}\big)dr(y)$$
$$= \big(e^{f(y)} - e^{g(y)}\big)dr(y)$$

누적 생성함수$\kappa(\cdot)$ 정의에 의하여 다음 수식을 만족한다.

$$\ln\int_y \big(e^{f(y)} - e^{g(y)}\big)\,dr(y) = \kappa[f] - \kappa[g] \leq \epsilon$$

$\int_Y \frac{g(y)}{dr}[dp(y) - dq(y)]dr = \epsilon\int_Y \frac{g(y)}{dr}dr \leq \epsilon'$ QED.

Ⅲ. 따름 정리 요약 증명

따름 정리 2(칸트로비치 거리와 KL-발산에서 정보량 관계)[6]: f = ▽D[p, r]와 g = ▽D[q, r]와 같은 기준 측도 r ∈ P (X)이 존재한다면, KL-발산 정보 $I(X,Y)$와 이에 해당하는 칸트로비치 정보(거리) $K_c[\omega, p \otimes q]$의 차이의 하한은 다음 수식을 만족한다.

$$K_c[\omega, p \otimes q] - I(X,Y)$$
$$\geq (\kappa[f] - \kappa[g])$$
$$+ \int_X g(x)\,[dp(x) - dq(x)]$$

여기서, $\kappa(\cdot) = \ln\int_Y e^{(\cdot)}\,dr(y)$는 정규화 상수이다.

요약 증명: 주변 측도 πXw = p 및 πYw = q 와 함께 결합 확률 측도 w ∈ P(X×Y)는 P(X×Y)에서 삼각형(w, p⊗q, p⊗p)을 정의한다고 하자. 칸트로비치 거리는 삼각부등식에 의해 다음 식을 만족한다[3].

$$K_c[\omega, p \otimes p] \leq K_c[\omega, p \otimes q] + K_c[p \otimes q, p \otimes p]$$
$$= K_c[\omega, p \otimes q] + K_c[q, p]$$

위 수식을 주변 측도 $\pi_X w$ = p 와 $\pi_Y w$ = q 의 곱 p⊗q ∈ P(X×Y)에 대한 결합 확률 측도 ω ∈ P(X×Y)의 칸트로비치 거리는 다음 식으로 나낼 수 있다.

$$K_c[\omega, p \otimes q] \geq K_c[\omega, p \otimes p] - K_c[q, p]$$

KL-발산은 칸트로비치 거리에서 주변 측도에 대한 제약이므로 $K_c[\omega, p \otimes p] \geq D[\omega, p \otimes p]$가 성립하므로[6] 다음 식으로 쉽게 표현될 수 있다.

$$K_c[\omega, p \otimes p] - K_c[q, p] \geq D[\omega, p \otimes p] - K_c[q, p]$$

가정 f = ▽D[p, r]과 g = ▽D[q, r]는 라그랑지 승수가 α = β = 1 인 것을 의미하며, 확률 측도는 p = exp(f-κ[f])r 와 q = exp(g-κ[g])r 의 형태를 갖고, 따름 정리 2 에 의해 다음 식이 된다.

$$D[\omega, p \otimes q] - K_c[q, p]$$
$$= D[\omega, p \otimes p] - D[q, p] + (\kappa[f] - \kappa[g])$$
$$+ \int_X g(x)\,[dp(x) - dq(x)]$$

이때 상호 정보는 $I(X,Y) = D[\omega, p \otimes p] - D[q, p]$ 이므로 다음 식을 만족한다.

$K_c[\omega, p \otimes q] \geq I(X,Y) + (\kappa[f] - \kappa[g]) + \int_X g(x)\,[dp(x) - dp(x)]$. 그러므로 정보량차이의 하한 값이 유도된다.QED.

ACKNOWLEDGMENT

본 논문은 한국전자통신연구원 내부연구개발사업 인공지능 기반 시맨틱 통신 기술 선행 연구(23RH1300, 23YH1500 매칭) 개발과제의 논문이다.

참 고 문 헌

[1] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” arXiv preprint arXiv:1701.07875, 2017.

[2] A. Korotin, A. Kolesov, and E. Burnaev, “Kantorovich Strikes Back! Wasserstein GANs are not Optimal Transport?” Advances in Neural Information Processing Systems, 2022.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” In Advances in neural information processing systems, pp. 2672-2680, 2014.

[4] C. Villani, “Optimal Transport: Old and New,” Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.

[5] L.V. Kantorovich, “On translocation of masses,” USSR AS Doklady 37(7-8), pp. 227-229, 1942 (in Russian). English translation: J. Math. Sci., vol. 133, no. 4, pp. 1381-1382, 2006.

[6] 이우용, 김경표, 김용선, “칸트로비치-바서스테인 거리 관점에서 샤논의 상호 정보량에 관한 연구,” CEIC 2019, Dec. 2019.