A Primal-Dual link between GANs and Autoencoders

Hisham Husain

Richard Nock Robert C. Williamson

Research School of Computer Science
CSIRO Data61

Australian National University

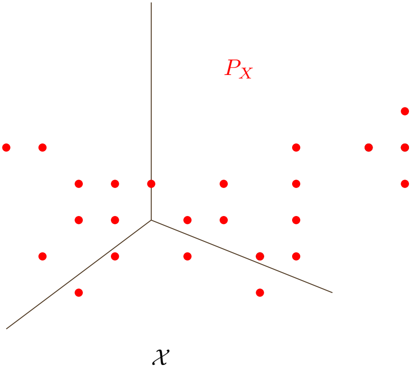
Outline

- Generative Models
- GANs and Autoencoders
- A Primal-Dual Relationship
- Implications / Conclusion

Generative Models

- Input space $\mathcal{X}$
- Target distribution P_X over $\mathcal{X}$

Generative Models



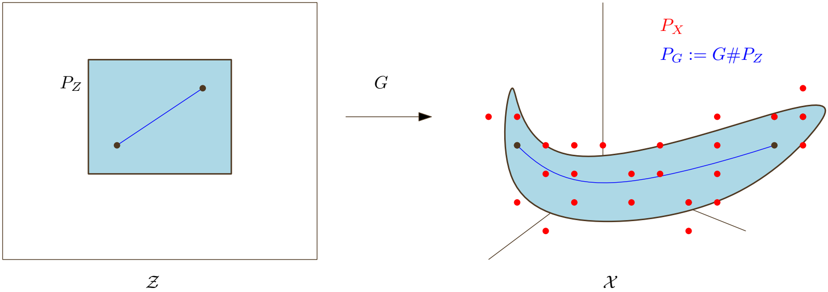
Generative Models

- Input space $\mathcal{X}$
- Target distribution P_X
- Latent space $\mathcal{Z}$
- Prior P_Z over $\mathcal{Z}$
- Generator $G : \mathcal{Z} \to \mathcal{X}$
- Model distribution $P_G = G\#P_Z$

Generative Models

- Input space $\mathcal{X}$
- Target distribution P_X
- Latent space $\mathcal{Z}$
- Prior P_Z over $\mathcal{Z}$
- Generator $G : \mathcal{Z} \to \mathcal{X}$
- Model distribution $P_G = G\#P_Z$
  If $X = G(z)$ where $z \sim P_Z$ then $X \sim G\#P_Z$.

Generative Models



Generative Models

- Take some discrepancy $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_{\geqslant 0}$

Generative Models

- Take some discrepancy $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_{\geqslant 0}$
- Find G that minimizes $D(P_G, P_X) = D(G\#P_Z, P_X)$

Generative Models

- Take some discrepancy $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_{\geqslant 0}$
- Find G that minimizes $D(P_G, P_X) = D(G\#P_Z, P_X)$
- How do we pick D?

CANDIDATES FOR D: f-DIVERGENCE

[Ali and Silvey, 1966]
For a convex function $f : \mathbb{R} \to (-\infty, \infty], f(1) = 0$

$$D(G\#P_Z, P_X) = D_f(P_X, P_G) = \int_{\mathcal{X}} f\left(\frac{dP_X}{dP_G}\right) dP_G \qquad (1)$$

CANDIDATES FOR D: INTEGRAL PROBABILITY METRIC

[Sriperumbudur et al., 2009]
For a function class $H \subseteq \mathcal{F}(\mathcal{X}, \mathbb{R})$,

$$D(G\#P_Z, P_X) = \mathrm{IPM}_H(G\#P_Z, P_X) \quad (2)$$

$$= \sup_{h \in H} \left\{ \int_{\mathcal{X}} h(x) dP_X(x) - \int_{\mathcal{X}} h(x) dP_G(x) \right\} \quad (3)$$

Candidates for D: Wasserstein Distance

[Villani, 2008]

- For some cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geqslant 0}$
- $\Pi(P_X, P_G) =$ $\left\{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) : \int_{\mathcal{X}} \pi(x, y)\, dy = P_X(x), \int_{\mathcal{X}} \pi(x, y)\, dx = P_G(y)\right\}$
- Wasserstein distance between P_X and P_G is

$$W_c(P_X, P_G) = \inf_{\pi \in \Pi(P_X, P_G)} \left\{ \int_{\mathcal{X} \times \mathcal{X}} c(x, y)\, d\pi(x, y) \right\} \quad (4)$$

Generative Adversarial Networks

[Goodfellow et al., 2014]

- Pick a set of discriminators $\mathcal{D} \subseteq \mathcal{F}(\mathcal{X}, (0,1))$.

$$D(G\#P_Z, P_X) = \sup_{d \in \mathcal{D}} \left\{ \mathbb{E}_{x \sim P_X}[\log(d(x))] + \mathbb{E}_{x \sim P_G}[\log(1 - d(x))] \right\} \tag{5}$$

Generative Adversarial Networks

[Goodfellow et al., 2014, Arjovsky et al., 2017]

- Pick a set of discriminators $\mathcal{D} \subseteq \mathcal{F}(\mathcal{X}, \mathbb{R})$.

$$D(G\#P_Z, P_X) = \sup_{d \in \mathcal{D}} \left\{ \mathbb{E}_{x \sim P_X}[\log(d(x))] + \mathbb{E}_{x \sim P_G}[\log(1 - d(x))] \right\} \tag{6}$$

$$D(G\#P_Z, P_X) = \sup_{d \in \mathcal{H}_c} \left\{ \mathbb{E}_{x \sim P_X}[d(x)] - \mathbb{E}_{x \sim P_G}[d(x)] \right\} \tag{7}$$

$$D(G\#P_Z, P_X) = \sup_{d \in \mathcal{D}} \left\{ \mathbb{E}_{x \sim P_X}[d(x)] - \mathbb{E}_{x \sim P_G}[f^\star(d(x))] \right\} \tag{8}$$

$$\dots \tag{9}$$

Generative Adversarial Networks

f-GAN Objective [Nowozin et al., 2016]

- Pick a convex function $f : \mathbb{R} \to \mathbb{R}$ with $f(1) = 0$.
- Pick a set of discriminators $\mathcal{D} \subseteq \mathcal{F}(\mathcal{X}, \mathbb{R})$.

$$\boxed{\mathrm{GAN}_f(G; \mathcal{D}) := \sup_{d \in \mathcal{D}} \left\{ \mathbb{E}_{x \sim P_X}[d(x)] - \mathbb{E}_{x \sim P_G}[f^\star(d(x))] \right\}} \quad (10)$$

where $f^\star(x) = \sup_y \left\{ x \cdot y - f(y) \right\}$ is the Legendre-Fenchel conjugate.

Generative Adversarial Networks

f-GAN Objective [Nowozin et al., 2016]

- Pick a convex function $f : \mathbb{R} \to \mathbb{R}$ with $f(1) = 0$.
- Pick a set of discriminators $\mathcal{D} \subseteq \mathcal{F}(\mathcal{X}, \mathrm{Dom}(f^\star))$.

$$\boxed{\mathrm{GAN}_f(G; \mathcal{D}) := \sup_{d \in \mathcal{D}} \left\{ \mathbb{E}_{x \sim P_X}[d(x)] - \mathbb{E}_{x \sim P_G}[f^\star(d(x))] \right\}} \tag{11}$$

where $f^\star(x) = \sup_y \left\{ x \cdot y - f(y) \right\}$ is the Legendre-Fenchel conjugate.

- If $\mathcal{D} = \mathcal{F}(\mathcal{X}, \mathrm{Dom}(f^\star))$, then $\mathrm{GAN}_f(G; D) = D_f(P_X, P_G)$ [Nguyen et al., 2010].

Autoencoders

- Encoder functions $E : \mathcal{X} \to \mathcal{P}(\mathcal{Z})$
- $E \in \mathcal{F}(\mathcal{X}, \mathcal{P}(\mathcal{Z}))$
- Reconstructing $x \in \mathcal{X}$ means

$$\mathbb{E}_{z \sim E(x)}[c(x, G(z))] \tag{12}$$

AUTOENCODERS

[Kingma and Welling, 2013, Zhao et al., 2017]

- Pick a reconstruction cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geqslant 0}$
- Pick a regularization function $\Omega : \mathcal{F}(\mathcal{X}, \mathcal{P}(\mathcal{Z})) \to \mathbb{R}_{\geqslant 0}$

$$D(G\#P_Z, P_X) = \tag{13}$$

$$\inf_{E \in \mathcal{F}(\mathcal{X}, \mathcal{P}(Z))} \left\{ \int_{\mathcal{X}} \mathbb{E}_{z \sim E(x)}[c(x, G(z))] dP_X(x) + \Omega(E) \right\} \tag{14}$$

Autoencoders

[Kingma and Welling, 2013, Zhao et al., 2017]

- Pick a reconstruction cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geqslant 0}$
- Pick a regularization function $\Omega : \mathcal{F}(\mathcal{X}, \mathcal{P}(\mathcal{Z})) \to \mathbb{R}_{\geqslant 0}$

$$D(G\#P_Z, P_X) = \tag{15}$$

$$\inf_{E \in \mathcal{F}(\mathcal{X}, \mathcal{P}(Z))} \left\{ \int_{\mathcal{X}} \mathbb{E}_{z \sim E(x)}[c(x, G(z))] \, dP_X(x) + \Omega(E) \right\} \tag{16}$$

- Autoencoder $\Omega(E) = 0$

Autoencoders

[Kingma and Welling, 2013, Zhao et al., 2017]

- Pick a reconstruction cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geqslant 0}$
- Pick a regularization function $\Omega : \mathcal{F}(\mathcal{X}, \mathcal{P}(\mathcal{Z})) \to \mathbb{R}_{\geqslant 0}$

$$D(G\#P_Z, P_X) = \tag{17}$$

$$\inf_{E \in \mathcal{F}(\mathcal{X}, \mathcal{P}(Z))} \left\{ \int_{\mathcal{X}} \mathbb{E}_{z \sim E(x)}[c(x, G(z))] dP_X(x) + \Omega(E) \right\} \tag{18}$$

- Autoencoder $\Omega(E) = 0$
- Variational Autoencoder $\Omega(E) = \int_{\mathcal{X}} \mathrm{KL}(E(x), P_Z) dP_X(x)$

Autoencoders

[Kingma and Welling, 2013, Zhao et al., 2017]

- Pick a reconstruction cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geqslant 0}$
- Pick a regularization function $\Omega : \mathcal{F}(\mathcal{X}, \mathcal{P}(\mathcal{Z})) \to \mathbb{R}_{\geqslant 0}$

$$D(G\#P_Z, P_X) = \tag{19}$$

$$\inf_{E \in \mathcal{F}(\mathcal{X}, \mathcal{P}(Z))} \left\{ \int_{\mathcal{X}} \mathbb{E}_{z \sim E(x)}[c(x, G(z))] dP_X(x) + \Omega(E) \right\} \tag{20}$$

- Autoencoder $\Omega(E) = 0$
- Variational Autoencoder $\Omega(E) = \int_{\mathcal{X}} \mathrm{KL}(E(x), P_Z) dP_X(x)$
- InfoVAE $\Omega(E) = \mathrm{KL}(E\#P_X, P_Z)$

Autoencoders

WAE Objective [Tolstikhin et al., 2017]

- Pick a reconstruction cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geqslant 0}$
- Pick a regularization function $\Omega : \mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z}) \to \mathbb{R}_{\geqslant 0}$

$$\mathrm{WAE}_{c,\lambda\cdot\Omega}(G) = \tag{21}$$

$$\boxed{\inf_{E\in\mathcal{F}(\mathcal{X},\mathcal{P}(Z))} \left\{ \int_{\mathcal{X}} \mathbb{E}_{z\sim E(x)}[c(x, G(z))]\, dP_X(x) + \lambda\Omega(E\#P_X, P_Z) \right\}} \tag{22}$$

- $E\#P_X$ is also referred to as the aggregated posterior.

Primal-Dual Optimization

- What is Duality?

Primal-Dual Optimization

- What is Duality?
- $\inf$ and $\sup$

Primal-Dual Optimization

- What is Duality?
- $\inf$ and $\sup$
- $\inf_a F(a) \geqslant \sup_b G(b)$ (Weak duality)

Primal-Dual Optimization

- What is Duality?
- $\inf$ and $\sup$
- $\inf_a F(a) \geqslant \sup_b G(a)$ (Weak duality)
- $\inf_a F(a) = \sup_b G(b)$ (Strong duality)

Primal-Dual Link

Theorem 1

Suppose $(\mathcal{X}, c)$ *is a metric space and let* $\mathcal{H}_c \subseteq \mathcal{F}(\mathcal{X}, \mathbb{R})$ *denote the set of* 1*-Lipschitz functions with respect to the metric* c*. Let* $f : \mathbb{R} \to (-\infty, \infty]$ *be a convex function with* $f(1) = 0$*. We have for any* $G : \mathcal{Z} \to \mathcal{X}$ *and all* $\lambda > 0$

$$\mathrm{WAE}_{c,\lambda \cdot D_f}(G) \geqslant \mathrm{GAN}_{\lambda f}(G; \mathcal{H}_c) \tag{23}$$

- Equality holds when G is invertible.
- Equality holds for *any* G if $\lambda > \lambda^*(P_X)$ for some finite $\lambda^*(P_X)$.
- Setting $f(x) = 1_{\{1\}}(x)$ with $G = \mathrm{Id}$ and $\mathcal{Z} = \mathcal{X}$ recovers the Kantorovich-Rubinstein duality.

Primal-Dual Link

Theorem 2

Suppose $(\mathcal{X}, c)$ *is a metric space and let* $\mathcal{H}_c \subseteq \mathcal{F}(\mathcal{X}, \mathbb{R})$ *denote the set of* 1*-Lipschitz functions with respect to the metric* c*. Let* $f : \mathbb{R} \to (-\infty, \infty]$ *be a convex function with* $f(1) = 0$*. We have for any* $G : \mathcal{Z} \to \mathcal{X}$ *and all* $\lambda > \lambda^*(P_X)$

$$\mathrm{WAE}_{c,\lambda \cdot D_f}(G) = \mathrm{GAN}_{\lambda f}(G; \mathcal{H}_c) = W_c(P_X, P_G) \tag{24}$$

Duality

Legendre-Fenchel Duality

$$D_f(P_X, P_G) = \sup_{d \in \mathcal{F}(\mathcal{X}, \mathbb{R})} \left\{ \mathbb{E}_{x \sim P_X}[d(x)] - \mathbb{E}_{x \sim P_G}[f^\star(d(x))] \right\}$$

Main Theorem

$$\mathrm{WAE}_{c, \lambda \cdot D_f}(G) = \sup_{d \in \mathcal{H}_c} \left\{ \mathbb{E}_{x \sim P_X}[d(x)] - \mathbb{E}_{x \sim P_G}[f^\star(d(x))] \right\}$$

Kantorovich-Rubenstein Duality

$$W_c(P_X, P_G) = \sup_{d \in \mathcal{H}_c} \left\{ \mathbb{E}_{x \sim P_X}[d(x)] - \mathbb{E}_{x \sim P_G}[d(x)] \right\}$$

Theoretical Application: Generalization Bound

- How do generative models generalize?

Theoretical Application: Generalization Bound

- How do generative models generalize?
- For GANs, it depends on the discriminator set $\mathcal{D}$. [Zhang et al., 2017]

Theoretical Application: Generalization Bound

- How do generative models generalize?
- For GANs, it depends on the discriminator set $\mathcal{D}$.
  [Zhang et al., 2017].
- What about Autoencoders?

Theoretical Application: Generalization Bound

- How do generative models generalize?
- For GANs, it depends on the discriminator set $\mathcal{D}$ [Zhang et al., 2017].
- What about Autoencoders?
- Apply duality with $\mathcal{D} = \mathcal{H}_c$

Theoretical Application: Generalization Bound

Theorem 3

Let $\widehat{\mathrm{WAE}}_{c,\lambda\cdot D_f}$ *denote the* $\mathrm{WAE}_{c,\lambda\cdot D_f}$ *objectives with* n *i.i.d samples for* P_X*. Assume that* $\Delta := \sup_{x,x'\in \mathrm{supp}(P_X)} c(x,x') < \infty$ *and suppose* S *is the* 1*-Upper Wasserstein dimension of* P_X *then we have*

$$\mathrm{WAE}_{c,\lambda\cdot D_f} \leqslant \widehat{\mathrm{WAE}}_{c,\lambda\cdot D_f} + O\left(n^{-1/S} + \Delta\sqrt{\frac{1}{n}\ln\left(\frac{1}{\delta}\right)}\right), \quad (25)$$

with probability at least $1-\delta$*.*

References I

- Ali, S. M. and Silvey, S. D. (1966).
  A general class of coefficients of divergence of one distribution from another.
  Journal of the Royal Statistical Society: Series B (Methodological), 28(1):131–142.

- Arjovsky, M., Chintala, S., and Bottou, L. (2017).
  Wasserstein gan.
  arXiv preprint arXiv:1701.07875.

- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014).
  Generative adversarial nets.
  In *Advances in neural information processing systems*, pages 2672–2680.

References II

Kingma, D. P. and Welling, M. (2013).
Auto-encoding variational bayes.
arXiv preprint arXiv:1312.6114.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010).
Estimating divergence functionals and the likelihood ratio by convex risk minimization.
IEEE Transactions on Information Theory, 56(11):5847–5861.

Nowozin, S., Cseke, B., and Tomioka, R. (2016).
f-gan: Training generative neural samplers using variational divergence minimization.
In *Advances in Neural Information Processing Systems*, pages 271–279.

References III

- Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. (2009).
  On integral probability metrics,\phi-divergences and binary classification.
  arXiv preprint arXiv:0901.2698.

- Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. (2017).
  Wasserstein auto-encoders.
  arXiv preprint arXiv:1711.01558.

- Villani, C. (2008).
  Optimal transport: old and new, volume 338.
  Springer Science & Business Media.

References IV

Zhang, P., Liu, Q., Zhou, D., Xu, T., and He, X. (2017).
On the discrimination-generalization tradeoff in gans.
arXiv preprint arXiv:1711.02771.

Zhao, S., Song, J., and Ermon, S. (2017).
Infovae: Information maximizing variational autoencoders.
arXiv preprint arXiv:1706.02262.