15 Generative Adversarial Networks (GAN)

Goal

Push-forward, Generative Adversarial Networks, min-max optimization, duality.

Alert 15.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.
This note is likely to be updated again soon.

Example 15.2: Simulating distributions

Suppose we want to sample from a Gaussian distribution with mean $\mathbf{u}$ and covariance S. The typical approach is to first sample from the standard Gaussian distribution (with zero mean and identity covariance) and then perform the transformation:

$$\text{If } \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbb{I}), \quad \text{then } X = \mathsf{T}(\mathbf{Z}) := \mathbf{u} + S^{1/2}\mathbf{Z} \sim \mathcal{N}(\mathbf{u}, S).$$

Similarly, we can sample from a χ^2 distribution with zero mean and degree d by the transformation:

$$\text{If } \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_d), \quad \text{then } X = \mathsf{T}(\mathbf{Z}) := \sum_{j=1}^{d} Z_j^2 \sim \chi^2(d).$$

In fact, we can sample from any distribution F on $\mathbb{R}$ by the following transformation:

$$\text{If } Z \sim \mathcal{N}(0,1), \quad \text{then } X = \mathsf{T}(Z) := \mathsf{F}^-(\Phi(Z)) \sim \mathsf{F}, \quad \text{where} \quad \mathsf{F}^-(z) = \min\{x : F(x) \geq z\},$$

and Φ is the cumulative distribution function of standard normal.

Theorem 15.3: Transforming to any probability measure

Let μ be a diffuse (Borel) probability measure on a polish space Z and similarly ν be any (Borel) probability measure on another polish space X. Then, there exist (measurable) maps $\mathsf{T} : \mathsf{Z} \to \mathsf{X}$ such that

$$\textit{If } Z \sim \mu, \quad \textit{then } X := \mathsf{T}(Z) \sim \nu.$$

■

Recall that a (Borel) probability measure is diffuse iff any single point has measure 0. For less mathematical readers, think of $\mathsf{Z} = \mathbb{R}^p$, $\mathsf{X} = \mathbb{R}^d$, μ and ν as probability densities on the respective Euclidean spaces.

Definition 15.4: Push-forward generative modeling

Given an i.i.d. sample $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim \chi$, we can now estimate the target density χ by the following push-forward approach:

$$\inf_{\boldsymbol{\theta}} \quad \mathsf{D}(\mathbf{X}, \mathsf{T}_{\boldsymbol{\theta}}(\mathbf{Z})),$$

where say $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_p)$, $\mathsf{T}_{\boldsymbol{\theta}} : \mathbb{R}^p \to \mathbb{R}^d$, and $\mathbf{X} \sim \chi$ (the true underlying data generating distribution). The function D is a "distance" that measures the closeness of our (true) data distribution (represented by $\mathbf{X}$) and model distribution (represented by $\mathsf{T}_{\boldsymbol{\theta}}(\mathbf{Z})$). By minimizing D we bring our model $\mathsf{T}_{\boldsymbol{\theta}}(\mathbf{Z})$ close to our data $\mathbf{X}$.

Remark 15.5: The good, the bad, and the beautiful

One big advantage of the push-forward approach in Definition 15.4 is that after training (e.g. finding a reasonable $\boldsymbol{\theta}$) we can *effortlessly* generate new data: we sample $\mathbf{Z} \in \mathcal{N}(\mathbf{0}, \mathbb{I}_d)$ and then set $\mathbf{X} = \mathsf{T}_{\boldsymbol{\theta}}(\mathbf{Z})$. On the flip side, we no longer have any explicit form for the model density (namely, that of $\mathsf{T}_{\boldsymbol{\theta}}(\mathbf{Z})$ when $p < d$). This renders direct maximum likelihood estimation of $\boldsymbol{\theta}$ impossible.

This is where we need the beautiful idea called duality. Basically, we need to distinguish two distributions: the data distribution represented by a sample $\mathbf{X}$ and the model distribution represented by a sample $\mathsf{T}_{\boldsymbol{\theta}}(\mathbf{Z})$. We distinguish them by running many tests, represented by functions f:

$$\sup_{f \in \mathcal{F}} |\mathbb{E} f(\mathbf{X}) - \mathbb{E} f\big(\mathsf{T}_{\boldsymbol{\theta}}(\mathbf{Z})\big)|.$$

If the class of tests $\mathcal{F}$ we run is dense enough, then we would be able to tell the difference between the two distributions and provide feedback for the model $\boldsymbol{\theta}$ to improve, until we no longer can tell the difference.

Definition 15.6: f-divergence (Csiszár 1963; Ali and Silvey 1966)

Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a strictly convex function (see the background lecture on optimization) with $f(1) = 0$. We define the following f-divergence to measure the closeness of two pdfs p and q:

$$\mathsf{D}_f(\mathsf{p} \| \mathsf{q}) := \int f\big(\mathsf{p}(\mathbf{x})/\mathsf{q}(\mathbf{x})\big) \cdot \mathsf{q}(\mathbf{x}) \, \mathrm{d}\mathbf{x}, \tag{15.1}$$

where we assume $\mathsf{q}(\mathbf{x}) = 0 \implies \mathsf{p}(\mathbf{x}) = 0$ (otherwise we put the divergence to ∞).

For two random variables $Z \sim \mathsf{q}$ and $X \sim \mathsf{p}$, we sometimes abuse the notation to mean

$$\mathsf{D}_f(X \| Z) := \mathsf{D}_f(\mathsf{p} \| \mathsf{q}).$$

Csiszár, I. (1963). "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten". *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének közleményei*, vol. 8, pp. 85–108.

Ali, S. M. and S. D. Silvey (1966). "A General Class of Coefficients of Divergence of One Distribution from Another". *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 28, no. 1, pp. 131–142.

Exercise 15.7: Properties of f-divergence

Prove the following:

- $\mathsf{D}_f(\mathsf{p} \| \mathsf{q}) \geq 0$, with 0 attained iff $\mathsf{p} = \mathsf{q}$;
- $\mathsf{D}_{f+g} = \mathsf{D}_f + \mathsf{D}_g$ and $\mathsf{D}_{sf} = s\mathsf{D}_f$ for $s > 0$;
- Let $g(t) = f(t) + s(t-1)$ for any s. Then, $\mathsf{D}_g = \mathsf{D}_f$;
- If $\mathsf{p}(\mathbf{x}) = 0 \iff \mathsf{q}(\mathbf{x}) = 0$, then $\mathsf{D}_f(\mathsf{p} \| \mathsf{q}) = \mathsf{D}_{f^\diamond}(\mathsf{q} \| \mathsf{p})$, where $f^\diamond(t) := t \cdot f(1/t)$;
- $f^\diamond$ is (strictly) convex, $f^\diamond(1) = 0$ and $(f^\diamond)^\diamond = f$;

The second last result indicates that f-divergences are not usually symmetric. However, we can always symmetrize them by the transformation: $f \leftarrow f + f^\diamond$.

Example 15.8: KL and LK

Let $f(t) = t \log t$, then we obtain the Kullback-Leibler (KL) divergence:

$$\mathsf{KL}(\mathsf{p}\|\mathsf{q}) = \int \mathsf{p}(\mathbf{x}) \log(\mathsf{p}(\mathbf{x})/\mathsf{q}(\mathbf{x}))\, \mathrm{d}\mathbf{x}.$$

Reverse the inputs we obtain the reverse KL divergence (a.k.a. Burg's entropy, Burg 1975):

$$\mathsf{LK}(\mathsf{p}\|\mathsf{q}) := \mathsf{KL}(\mathsf{q}\|\mathsf{p}).$$

Verify by yourself that the underlying function $f = -\log$ for reverse KL.

Burg, J. P. (1975). "Maximum Entropy Spectral Analysis". PhD thesis.

Example 15.9: More divergences, more fun

Derive the formula for the following f-divergences:

- Pearson's χ^2-divergence: $f(t) = \frac{1}{2}(t-1)^2$;
- Neyman's reverse χ^2-divergence: $f(t) = \frac{1}{2}(\frac{1}{t} + t - 2) = \frac{t}{2}(\frac{1}{t} - 1)^2$;
- Hellinger divergence: $f(t) = 2(\sqrt{t} - 1)^2$;
- total variation: $f(t) = |t - 1|$;
- Jensen-Shannon divergence: $f(t) = t \log t - (t+1)\log(t+1) + \log 4$;
- Rényi divergence (Rényi 1961, 1965): $f_\alpha(t) = \frac{t^\alpha - 1 - \alpha(t-1)}{\alpha(\alpha-1)}$ for some $\alpha \in \mathbb{R}$. What are the limits for $\alpha = 1$ and $\alpha = 0$? Show that $f_\alpha^\diamond(t) = f_{1-\alpha}(t)$.

Which of the above are symmetric?

Rényi, A. (1961). "On Measures of Entropy and Information". In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 547–561.

— (1965). "On the Foundations of Information Theory". *Review of the International Statistical Institute*, vol. 33, no. 1, pp. 1–14.

Remark 15.10: More on the Rényi divergence

(Tsallis 1988; Perez 1967; Havrda and Charvát 1967)

Tsallis, C. (1988). "Possible generalization of Boltzmann-Gibbs statistics". *Journal of statistical physics*, vol. 52, no. 1, pp. 479–487.

Perez, A. (1967). "Information-theoretic risk estimates in statistical decision". *Kybernetika*, vol. 3, no. 1, pp. 1–21.

Havrda, J. and F. Charvát (1967). "Quantification method of classification processes. Concept of structural a-entropy". *Kybernetika*, vol. 3, no. 1, pp. 30–35.

Definition 15.11: Fenchel conjugate function

For any extended real-valued function $f : \mathsf{V} \to (-\infty, \infty]$ we define its Fenchel conjugate function as:

$$f^*(\mathbf{x}^*) := \sup_{\mathbf{x}} \langle \mathbf{x}, \mathbf{x}^* \rangle - f(\mathbf{x}).$$

We remark that f^* is always a convex function (of $\mathbf{x}^*$).

If $\operatorname{dom} f$ is nonempty and closed, and f is continuous, then

$$f^{**} := (f^*)^* = f.$$

This remarkable property of convex functions will now be used!

Example 15.12: Fenchel conjugate of JS

Consider the convex function that defines the Jensen-Shannon divergence:

$$f(t) = t \log t - (t+1)\log(t+1) + \log 4. \tag{15.2}$$

We derive its Fenchel conjugate:

$$f^*(s) = \sup_t st - f(t) = \sup_t st - t\log t + (t+1)\log(t+1) - \log 4.$$

Taking derivative w.r.t. t we obtain

$$s - \log t - 1 + \log(t+1) + 1 = 0 \iff t = \frac{1}{\exp(-s) - 1},$$

and plugging it back we get

$$\begin{aligned} f^*(s) &= \frac{s}{\exp(-s)-1} - \frac{1}{\exp(-s)-1}\log\frac{1}{\exp(-s)-1} + \frac{\exp(-s)}{\exp(-s)-1}\log\frac{\exp(-s)}{\exp(-s)-1} - \log 4 \\ &= \frac{s}{\exp(-s)-1} - \frac{1}{\exp(-s)-1}\log\frac{1}{\exp(-s)-1} + \frac{\exp(-s)}{\exp(-s)-1}\log\frac{1}{\exp(-s)-1} - \frac{s\exp(-s)}{\exp(-s)-1} - \log 4 \\ &= -s - \log(\exp(-s)-1) - \log 4 \\ &= -\log(1-\exp(s)) - \log 4. \end{aligned} \tag{15.3}$$

Using conjugation again, we obtain the important formula:

$$f(t) = \sup_s st - f^*(s) = \sup_s st + \log(1-\exp(s)) + \log 4.$$

Exercise 15.13: More conjugates

Derive the Fenchel conjugate of the other convex functions in Example 15.8 and Example 15.9.

Definition 15.14: Generative adversarial networks (GAN) (Goodfellow et al. 2014)

We are now ready to define the original GAN, which amounts to using the Jensen-Shannon divergence in Definition 15.4:

$$\inf_{\boldsymbol{\theta}} \quad \mathsf{JS}(\mathbf{X} \| \mathsf{T}_{\boldsymbol{\theta}}(\mathbf{Z})), \quad \text{where} \quad \mathsf{JS}(\mathsf{p}\|\mathsf{q}) = \mathsf{D}_f(\mathsf{p}\|\mathsf{q}) = \mathsf{KL}(\mathsf{p}\|\tfrac{\mathsf{p}+\mathsf{q}}{2}) + \mathsf{KL}(\mathsf{p}\|\tfrac{\mathsf{p}+\mathsf{q}}{2}),$$

and the convex function f is defined in (15.2), along with its Fenchel conjugate f^* given in (15.3).

To see how we can circumvent the lack of an explicit form of the density $\mathsf{q}(\mathbf{x})$ of $\mathsf{T}_{\boldsymbol{\theta}}(\mathbf{Z})$, we expand using duality:

$$\begin{aligned} \mathsf{JS}(\mathbf{X}\|\mathsf{T}_{\boldsymbol{\theta}}(\mathbf{Z})) &= \int_{\mathbf{x}} f\big(\mathsf{p}(\mathbf{x})/\mathsf{q}(\mathbf{x})\big)\mathsf{q}(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= \int_{\mathbf{x}} [\sup_s s\mathsf{p}(\mathbf{x})/\mathsf{q}(\mathbf{x}) - f^*(s)]\mathsf{q}(\mathbf{x})\,\mathrm{d}\mathbf{x} \end{aligned}$$

$$
\begin{aligned}
&= \int_{\mathbf{x}} [\sup_s s\mathsf{p}(\mathbf{x}) - f^*(s)\mathsf{q}(\mathbf{x})]\,\mathrm{d}\mathbf{x} \\
&= \sup_{\mathsf{S}:\mathbb{R}^d\to\mathbb{R}} \int_{\mathbf{x}} \mathsf{S}(\mathbf{x})\mathsf{p}(\mathbf{x})\,\mathrm{d}\mathbf{x} - \int_{\mathbf{x}} f^*(\mathsf{S}(\mathbf{x}))\mathsf{q}(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= \sup_{\mathsf{S}:\mathbb{R}^d\to\mathbb{R}} \mathbb{E}\mathsf{S}(\mathbf{X}) - \mathbb{E}f^*(\mathsf{S}(\mathsf{T}_{\boldsymbol{\theta}}(\mathbf{Z}))).
\end{aligned}
$$

Therefore, if we parameterize the test function S by $\boldsymbol{\phi}$ (say a deep net), then we obtain a lower bound of the Jensen-Shannon divergence for minimizing:

$$
\inf_{\boldsymbol{\theta}} \sup_{\boldsymbol{\phi}} \mathbb{E}\mathsf{S}_{\boldsymbol{\phi}}(\mathbf{X}) - \mathbb{E}f^*(\mathsf{S}_{\boldsymbol{\phi}}(\mathsf{T}_{\boldsymbol{\theta}}(\mathbf{Z}))).
$$

Of course, we cannot compute either of the two expectations, so we use sample average to approximate them:

$$
\inf_{\boldsymbol{\theta}} \sup_{\boldsymbol{\phi}} \hat{\mathbb{E}}\mathsf{S}_{\boldsymbol{\phi}}(\mathbf{X}) - \hat{\mathbb{E}}f^*(\mathsf{S}_{\boldsymbol{\phi}}(\mathsf{T}_{\boldsymbol{\theta}}(\mathbf{Z}))), \tag{15.4}
$$

where the first sample expectation $\hat{\mathbb{E}}$ is simply the average of the given training data while the second sample expectation is the average over samples generated by the model $\mathsf{T}_{\boldsymbol{\theta}}(\mathbf{Z})$ (recall Remark 15.5).

In practice, both $\mathsf{T}_{\boldsymbol{\theta}}$ and $\mathsf{S}_{\boldsymbol{\phi}}$ are represented by deep nets, and the former is called the generator while the latter is called the discriminator. Our final objective (15.4) represents a two-player game between the generator and the discriminator. At equilibrium (if any) the generator is forced to mimic the (true) data distribution (otherwise the discriminator would be able to tell the difference and incur a loss for the generator).

See the background lecture on optimization for a simple algorithm (gradient-descent-ascent) for solving (15.4).

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*.

Remark 15.15: Approximation

We made a number of approximations in Definition 15.14. Thus, technically speaking, the final GAN objective (15.4) no longer minimizes the Jensen-Shannon divergence. Nock et al. (2017) and Liu et al. (2017) formally studied this approximation trade-off.

Nock, R., Z. Cranko, A. K. Menon, L. Qu, and R. C. Williamson (2017). "f-GANs in an Information Geometric Nutshell". In: *Advances in Neural Information Processing Systems*.

Liu, S., L. Bottou, and K. Chaudhuri (2017). "Approximation and convergence properties of generative adversarial learning". In: *Advances in Neural Information Processing Systems*.

Exercise 15.16: Catch me if you can

Let us consider the game between the generator $\mathsf{q}(\mathbf{x})$ (the implicit density of $\mathsf{T}_{\boldsymbol{\theta}}(\mathbf{Z})$) and the discriminator $\mathsf{S}(\mathbf{x})$:

$$
\inf_{\mathsf{q}} \sup_{\mathsf{S}} \int_{\mathbf{x}} \mathsf{S}(\mathbf{x})\mathsf{p}(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int_{\mathbf{x}} \log\big(1 - \exp(\mathsf{S}(\mathbf{x}))\big)\mathsf{q}(\mathbf{x})\,\mathrm{d}\mathbf{x} + \log 4.
$$

- Fixing the generator q, what is the optimal discriminator S?
- Plugging the optimal discriminator S back in, what is the optimal generator?
- Fixing the discriminator S, what is the optimal generator q?
- Plugging the optimal generator q back in, what is the optimal discriminator?

Exercise 15.17: KL vs. LK

Recall that the f-divergence $\mathsf{D}_f(\mathsf{p}\|\mathsf{q})$ is infinite iff for some $\mathbf{x}$, $\mathsf{p}(\mathbf{x}) \neq 0$ while $\mathsf{q}(\mathbf{x}) = 0$. Consider the following twin problems:

$$\mathsf{q}_{\mathsf{KL}} := \operatorname*{argmin}_{\mathsf{q}\in\mathcal{Q}} \ \mathsf{KL}(\mathsf{p}\|\mathsf{q})$$
$$\mathsf{q}_{\mathsf{LK}} := \operatorname*{argmin}_{\mathsf{q}\in\mathcal{Q}} \ \mathsf{LK}(\mathsf{p}\|\mathsf{q}).$$

Recall that $\operatorname{supp}(\mathsf{p}) := \operatorname{cl}\{\mathbf{x} : \mathsf{p}(\mathbf{x}) \neq \mathbf{0}\}$. What can we say about $\operatorname{supp}(\mathsf{p})$, $\operatorname{supp}(\mathsf{q}_{\mathsf{KL}})$ and $\operatorname{supp}(\mathsf{q}_{\mathsf{LK}})$? What about JS?

Definition 15.18: KL GAN (Sønderby et al. 2017)

Sønderby, C. K., J. Caballero, L. Theis, W. Shi, and F. Huszár (2017). "Amortised MAP Inference for Image Super-resolution". In: *International Conference on Learning Representations*.

Definition 15.19: f-GAN (Nowozin et al. 2016)

Following Nowozin et al. (2016), we summarize the main idea of f-GAN as follows:

- Generator: Let μ be a fixed reference probability measure on space Z (usually the standard normal distribution) and $Z \sim \mu$. Let ν be any target probability measure on space X and $X \sim \nu$. Let $\mathcal{T} \subseteq \{\mathsf{T} : \mathsf{Z} \to \mathsf{X}\}$ be a class of transformations. According to Theorem 15.3 we know there exist transformations T (which *may or may not* be in our class $\mathcal{T}$) so that $\mathsf{T}(Z) \sim X \sim \nu$. Our goal is to approximate such transformations T using our class $\mathcal{T}$.
- Loss: We use the f-divergence to measure the closeness between the target X and the transformed reference $\mathsf{T}(Z)$:

$$\inf_{\mathsf{T}\in\mathcal{T}} \ \mathsf{D}_f\big(X\|\mathsf{T}(Z)\big).$$

  In fact, any loss function that allows us to distinguish two probability measures can be used. However, we face an additional difficulty here: the densities of X and $\mathsf{T}(Z)$ (w.r.t. a third probability measure λ) are not known to us (especially the former) so we cannot naively evaluate the f-divergence in (15.1).
- Discriminator: A simple variational reformulation will resolve the above difficulty! Indeed,

$$\begin{aligned}
\mathsf{D}_f(X\|\mathsf{T}(Z)) &= \int f\left(\frac{\mathrm{d}\nu}{\mathrm{d}\tau}(\mathbf{x})\right)\mathrm{d}\tau(\mathbf{x}) && (\mathsf{T}(Z)\sim\tau)\\
&= \int \sup_{s\in\operatorname{dom}(f^*)}\left[s\frac{\mathrm{d}\nu}{\mathrm{d}\tau}(\mathbf{x}) - f^*(s)\right]\mathrm{d}\tau(\mathbf{x}) && (f^{**}=f)\\
&\geq \sup_{\mathsf{S}\in\mathcal{S}} \int\left[S(\mathbf{x})\frac{\mathrm{d}\nu}{\mathrm{d}\tau}(\mathbf{x}) - f^*(S(\mathbf{x}))\right]\mathrm{d}\tau(\mathbf{x}) && (\mathcal{S}\subseteq\{\mathsf{S}:\mathsf{X}\to\operatorname{dom}(f^*)\})\\
&= \sup_{\mathsf{S}\in\mathcal{S}} \mathbf{E}[\mathsf{S}(X)] - \mathbf{E}[f^*\big(\mathsf{S}(\mathsf{T}(Z))\big)] && (\text{equality if } f'\left(\frac{\mathrm{d}\nu}{\mathrm{d}\tau}\right)\in\mathcal{S}),
\end{aligned}$$

  so our estimation problem reduces to the following minimax zero-sum game:

$$\inf_{\mathsf{T}\in\mathcal{T}} \sup_{\mathsf{S}\in\mathcal{S}} \ \mathbf{E}[\mathsf{S}(X)] - \mathbf{E}[f^*\big(\mathsf{S}(\mathsf{T}(Z))\big)].$$

  By replacing the expectations with empirical averages we can (approximately) solve the above problem with classic stochastic algorithms.

- Reparameterization: The class of functions $\mathcal{S}$ we use to test the difference between two probability measures in the f-divergence must have their range contained in the domain of f^*. One convenient way to enforce this constraint is to set

$$\mathcal{S} = \sigma \circ \mathcal{U} := \{\sigma \circ \mathsf{U} : \mathsf{U} \in \mathcal{U}\}, \quad \sigma : \mathbb{R} \to \operatorname{dom}(f^*), \quad \mathcal{U} \subseteq \{\mathsf{U} : \mathsf{X} \to \mathbb{R}\},$$

where the functions U are unconstrained and the domain constraint is enforced through a *fixed* "activation function" σ. With this choice, the final f-GAN problem we need to solve is:

$$\inf_{\mathsf{T} \in \mathcal{T}} \sup_{\mathsf{U} \in \mathcal{U}} \mathbf{E}[\sigma \circ \mathsf{U}(X)] - \mathbf{E}[(f^* \circ \sigma)\big(\mathsf{U}(\mathsf{T}(Z))\big)].$$

Typically we choose an increasing σ so that the composition $f^* \circ \sigma$ is "nice." Note that the monotonicity of σ implies the same monotonicity of the composition $f^* \circ \sigma$ (since f^* is always increasing as f is defined only on $\mathbb{R}_+$). In this case, we prefer to pick a test function U so that $\mathsf{U}(X)$ is large while $\mathsf{U}(\mathsf{T}(Z))$ is small. This choice aligns with the goal to "maximize target and minimize transformed reference," although the opposite choice would work equally well (merely a sign change).

Nowozin, S., B. Cseke, and R. Tomioka (2016). "f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization". In: *Advances in Neural Information Processing Systems*.

Remark 15.20: f-GAN recap

To specify an f-GAN, we need:

- A reference probability measure μ: should be easy to sample and typically we use standard normal;
- A class of transformations (generators): $\mathcal{T} \subseteq \{\mathsf{T} : \mathsf{Z} \to \mathsf{X}\}$;
- An increasing convex function $f^* : \operatorname{dom}(f^*) \to \mathbb{R}$ with $f^*(0) = 0$ and $f^*(s) \geq s$ (or equivalently an f-divergence);
- An increasing activation function $\sigma : \mathbb{R} \to \operatorname{dom}(f^*)$ so that $f^* \circ \sigma$ is "nice";
- A class of *unconstrained* test functions (discriminators): $\mathcal{U} \subseteq \{\mathsf{U} : \mathsf{X} \to \mathbb{R}\}$ so that $\mathcal{S} = \sigma \circ \mathcal{U}$.

More on f-GAN: (Mescheder et al. 2018).

Mescheder, L., A. Geiger, and S. Nowozin (2018). "Which Training Methods for GANs do actually Converge?" In: *Proceedings of the 35th International Conference on Machine Learning*.

Definition 15.21: Wasserstein GAN (WGAN) (Arjovsky et al. 2017)

If we let the test functions range over the set of all 1-Lipschitz continuous functions $\mathcal{L}$, we then obtain WGAN:

$$\inf_{\boldsymbol{\theta}} \sup_{\mathsf{S} \in \mathcal{L}} \mathbb{E}\mathsf{S}(\mathbf{X}) - \mathbb{E}\mathsf{S}\big(\mathsf{T}_{\boldsymbol{\theta}}(\mathbf{Z})\big),$$

which corresponds to the dual of the 1-Wasserstein distance.

Gulrajani et al. 2017; Salimans et al. 2016; Petzka et al. 2018; Wei et al. 2018; Adler and Lunz 2018; Liu et al. 2018.

Arjovsky, M., S. Chintala, and L. Bottou (2017). "Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*.

Gulrajani, I., F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville (2017). "Improved Training of Wasserstein GANs". In: *Advances in Neural Information Processing Systems*.

Salimans, T., I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen (2016). "Improved Techniques for Training GANs". In: *Advances in Neural Information Processing Systems*.

Petzka, H., A. Fischer, and D. Lukovnikov (2018). "On the regularization of Wasserstein GANs". In: *International Conference on Learning Representations*.
Wei, X., B. Gong, Z. Liu, W. Lu, and L. Wang (2018). "Improving the Improved Training of Wasserstein GANs: A Consistency Term and Its Dual Effect". In: *International Conference on Learning Representations*.
Adler, J. and S. Lunz (2018). "Banach Wasserstein GAN". In: *Advances in Neural Information Processing Systems*.
Liu, H., X. GU, and D. Samaras (2018). "A Two-Step Computation of the Exact GAN Wasserstein Distance". In: *Proceedings of the 35th International Conference on Machine Learning*.

Definition 15.22: Maximum Mean Discrepancy GAN (MMD-GAN)

If, instead, we choose the test functions from a reproducing kernel Hilbert space (RKHS), then we obtain the so-called MMD-GAN (Dziugaite et al. 2015; Li et al. 2015; Li et al. 2017a; Bińkowski et al. 2018):

$$\inf_{\boldsymbol{\theta}} \sup_{\mathsf{S}\in\mathcal{H}_\kappa} \mathbb{E}\mathsf{S}(\mathbf{X}) - \mathbb{E}\mathsf{S}\big(\mathsf{T}_{\boldsymbol{\theta}}(\mathbf{Z})\big),$$

where $\mathcal{H}_\kappa$ is the unit ball of the RKHS induced by the kernel κ.

More on MMD-GAN: (Li et al. 2017a; Wang et al. 2019; Mroueh et al. 2017; Arbel et al. 2018; Dai et al. 2017; Li et al. 2018; Ren et al. 2016; Sutherland et al. 2017; Zhao et al. 2017; Li et al. 2017b).

Dziugaite, G. K., D. M. Roy, and Z. Ghahramani (2015). "Training generative neural networks via maximum mean discrepancy optimization". In: *Conference on Uncertainty in Artificial Intelligence*.
Li, Y., K. Swersky, and R. Zemel (2015). "Generative Moment Matching Networks". In: *Proceedings of the 32nd International Conference on Machine Learning*.
Li, C.-L., W.-C. Chang, Y. Cheng, Y. Yang, and B. Poczos (2017a). "MMD GAN: Towards Deeper Understanding of Moment Matching Network". In: *Advances in Neural Information Processing Systems*.
Bińkowski, M., D. J. Sutherland, M. Arbel, and A. Gretton (2018). "Demystifying MMD GANs". In: *International Conference on Learning Representations*.
Wang, W., Y. Sun, and S. Halgamuge (2019). "Improving MMD-GAN Training with Repulsive Loss Function". In: *International Conference on Learning Representations*.
Mroueh, Y., T. Sercu, and V. Goel (2017). "McGan: Mean and Covariance Feature Matching GAN". In: *Proceedings of the 34th International Conference on Machine Learning*.
Arbel, M., D. J. Sutherland, M. Bińkowski, and A. Gretton (2018). "On gradient regularizers for MMD GANs". In: *Advances in Neural Information Processing Systems*.
Dai, Z., A. Almahairi, P. Bachman, E. Hovy, and A. Courville (2017). "Calibrating Energy-based Generative Adversarial Networks". In: *International Conference on Learning Representations*.
Li, C.-L., M. Zaheer, Y. Zhang, B. Poczos, and R. Salakhutdinov (2018). "Point Cloud GAN". arXiv:1810.05795.
Ren, Y., J. Zhu, J. Li, and Y. Luo (2016). "Conditional Generative Moment-Matching Networks". In: *Advances in Neural Information Processing Systems*.
Sutherland, D. J., H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton (2017). "Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy". In: *International Conference on Learning Representations*.
Zhao, J., M. Mathieu, and Y. LeCun (2017). "Energy-based Generative Adversarial Networks". In: *International Conference on Learning Representations*.
Li, Y., A. Schwing, K.-C. Wang, and R. Zemel (2017b). "Dualing GANs". In: *Advances in Neural Information Processing Systems*.

Definition 15.23: Fisher GAN (Mroueh and Sercu 2017)

Mao et al. 2017; Tao et al. 2018; Bellemare et al. 2017

Mroueh, Y. and T. Sercu (2017). "Fisher GAN". In: *Advances in Neural Information Processing Systems*.
Mao, X., Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley (2017). "Least Squares Generative Adversarial Networks". In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 2813–2821.
Tao, C., L. Chen, R. Henao, J. Feng, and L. Carin (2018). "χ^2 Generative Adversarial Network". In: *Proceedings of the 35th International Conference on Machine Learning*.
Bellemare, M. G., I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos (2017). "The Cramer Distance as a Solution to Biased Wasserstein Gradients". arXiv:1705.10743.

Definition 15.24: Sobolev GAN (Mroueh et al. 2018)

Mroueh et al. 2019; Mroueh and Rigotti 2020

Mroueh, Y., C.-L. Li, T. Sercu, A. Raj, and Y. Cheng (2018). "Sobolev GAN". In: *International Conference on Learning Representations*.

Mroueh, Y., T. Sercu, and A. Raj (2019). "Sobolev Descent". In: *International Conference on Artificial Intelligence and Statistics*.

Mroueh, Y. and M. Rigotti (2020). "Unbalanced Sobolev Descent". In: *Advances in Neural Information Processing Systems*.