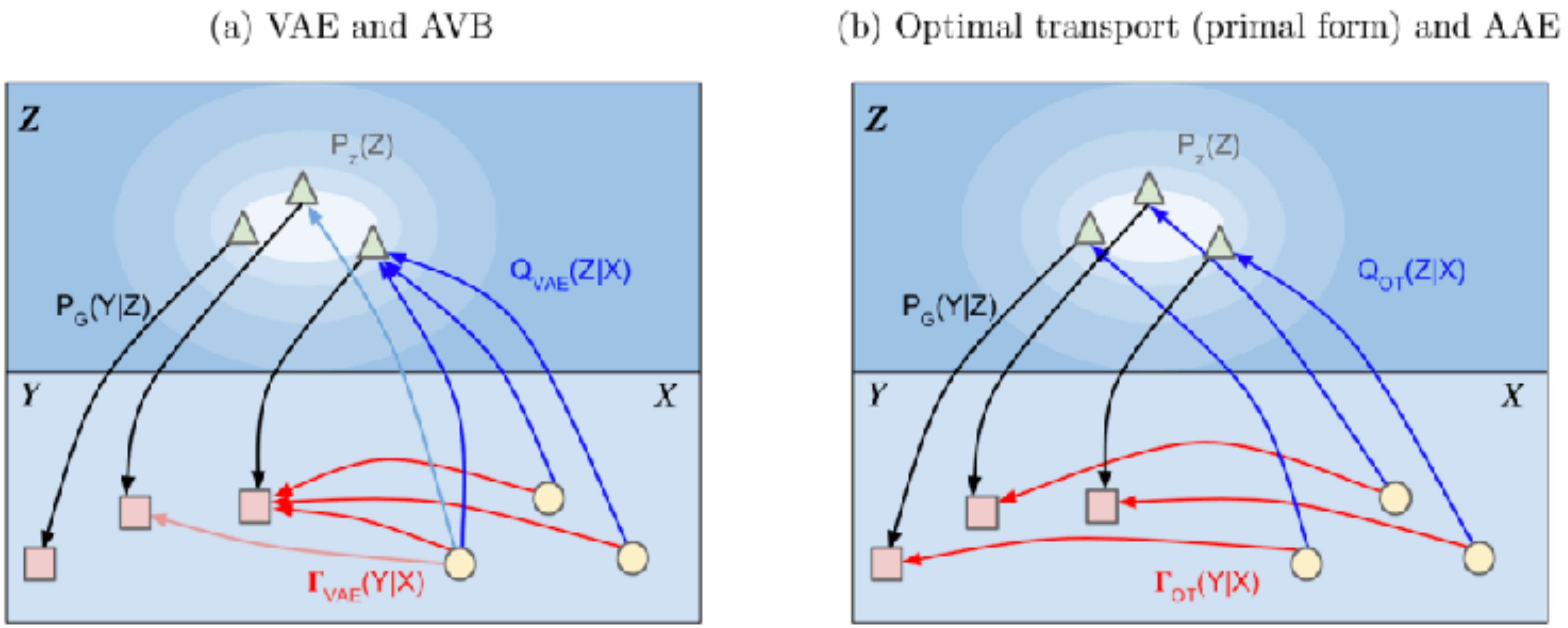
From OT to generative modeling: the VEGAN cookbook

Cheng Xin

Notations

- set: calligraphic letters, $\mathcal{X}$
- random variables: capital letters, X
- random variable values: lower case letters, x
- probability distributions: capital letters functions, $P(X)$
- probability density: lower case letters functions, $p(x)$


(a) VAE and AVB

(b) Optimal transport (primal form) and AAE

Variational auto-encoders (VAE) [2] utilize models P_G of the form (3) and minimize

$$D_{\mathrm{VAE}}(P_X, P_G) = \inf_{Q(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}\left[D_{\mathrm{KL}}\big(Q(Z|X), P_Z\big) - \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)]\right] \tag{5}$$

Adversarial variational Bayes (AVB) [6]

$$D_{\mathrm{AVB}}(P_X, P_G) = \inf_{Q_e(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}\left[D_{\mathrm{f,GAN}}\big(Q_e(Z|X), P_Z\big) - \mathbb{E}_{Q_e(Z|X)}[\log p_G(X|Z)]\right]. \tag{6}$$

Adversarial auto-encoders (AAE) [1] replace the D_{KL} term in (5) with another regularizer:

$$D_{\mathrm{AAE}}(P_X, P_G) = \inf_{Q(Z|X)\in\mathcal{Q}} D_{\mathrm{GAN}}(Q_Z, P_Z) - \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)], \tag{7}$$

Wasserstein Distance

$$W_c(P, Q) := \inf_{\Gamma \in \mathcal{P}(X \sim P, Y \sim Q)} \mathbb{E}_{(X,Y) \sim \Gamma}[c(X, Y)] \,, \tag{1}$$

Kantorovich-Rubinstein theorem

$$W_1(P, Q) = \sup_{f \in \mathcal{F}_L} \left| \mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)] \right|, \tag{2}$$

GAN and Wasserstein-GAN

$$D_{\mathrm{GAN}}(P_X, P_G) = \sup_{T \in \mathcal{T}} \mathbb{E}_{X \sim P_X}\left[\log T(X)\right] + \mathbb{E}_{Z \sim P_Z}\left[\log\left(1 - T(G(Z))\right)\right] \tag{4}$$

$$D_{\mathrm{WGAN}}(P_X, P_G) = \sup_{T \in \mathcal{W}} \mathbb{E}_{X \sim P_X}\left[T(X)\right] - \mathbb{E}_{Z \sim P_Z}\left[T(G(Z))\right],$$

Decomposition

Formal statement As in Section 2, $\mathcal{P}(X \sim P_X, Y \sim P_G)$ denotes the set of all joint distributions of (X, Y) with marginals P_X, P_G, and likewise for $\mathcal{P}(X \sim P_X, Z \sim P_Z)$. The set of all joint distributions of (X, Y, Z) such that $X \sim P_X$, $(Y, Z) \sim P_{G,Z}$, and $(Y \perp\!\!\!\perp X)|Z$ will be denoted by $\mathcal{P}_{X,Y,Z}$. Finally, we denote by $\mathcal{P}_{X,Y}$ and $\mathcal{P}_{X,Z}$ the sets of marginals on (X, Y) and (X, Z) (respectively) induced by distributions in $\mathcal{P}_{X,Y,Z}$. Note that $\mathcal{P}(P_X, P_G)$, $\mathcal{P}_{X,Y,Z}$, and $\mathcal{P}_{X,Y}$ depend on the choice of conditional distributions $P_G(Y|Z)$, while $\mathcal{P}_{X,Z}$ does not. In fact, it is easy to check that $\mathcal{P}_{X,Z} = \mathcal{P}(X \sim P_X, Z \sim P_Z)$. From the definitions it is clear that $\mathcal{P}_{X,Y} \subseteq \mathcal{P}(P_X, P_G)$ and we get the following upper bound:

$$W_c(P_X, P_G) \leq W_c^\dagger(P_X, P_G) := \inf_{P \in \mathcal{P}_{X,Y}} \mathbb{E}_{(X,Y)\sim P}\left[c(X, Y)\right] \tag{9}$$

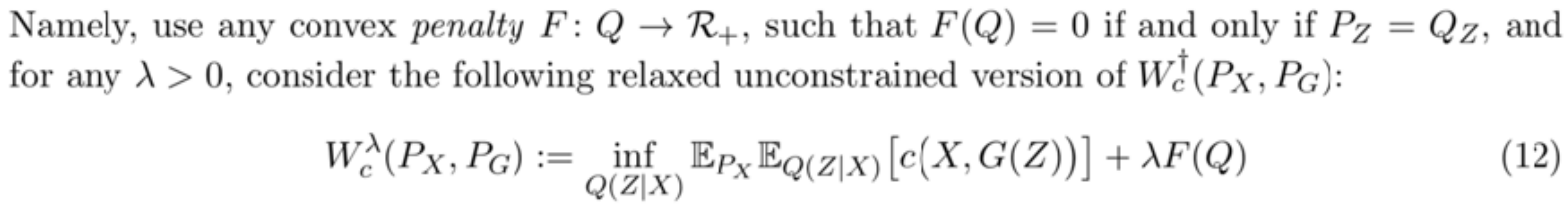
If $P_G(Y|Z)$ are Dirac measures (i.e., $Y = G(Z)$), the two sets are actually coincide, thus justifying the reparametrization (8) and the illustration in Figure 1(b), as demonstrated in the following theorem:

Theorem 1. *If $P_G(Y|Z = z) = \delta_{G(z)}$ for all $z \in \mathcal{Z}$, where $G\colon \mathcal{Z} \to \mathcal{X}$, we have*

$$W_c(P_X, P_G) = W_c^\dagger(P_X, P_G) = \inf_{P \in \mathcal{P}(X\sim P_X, Z\sim P_Z)} \mathbb{E}_{(X,Z)\sim P}\big[c\big(X, G(Z)\big)\big] \tag{10}$$

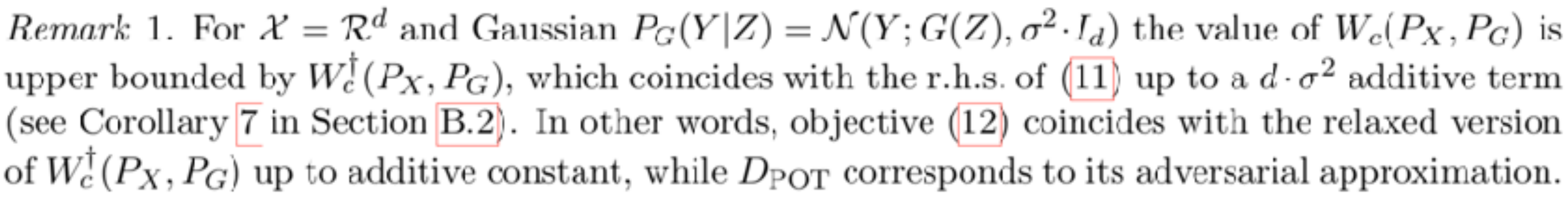
$$= \inf_{Q\colon Q_Z = P_Z} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}\big[c\big(X, G(Z)\big)\big], \tag{11}$$

$$W_c^\lambda(P_X, P_G) := \inf_{Q(Z|X)} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}\big[c\big(X, G(Z)\big)\big] + \lambda F(Q) \tag{12}$$

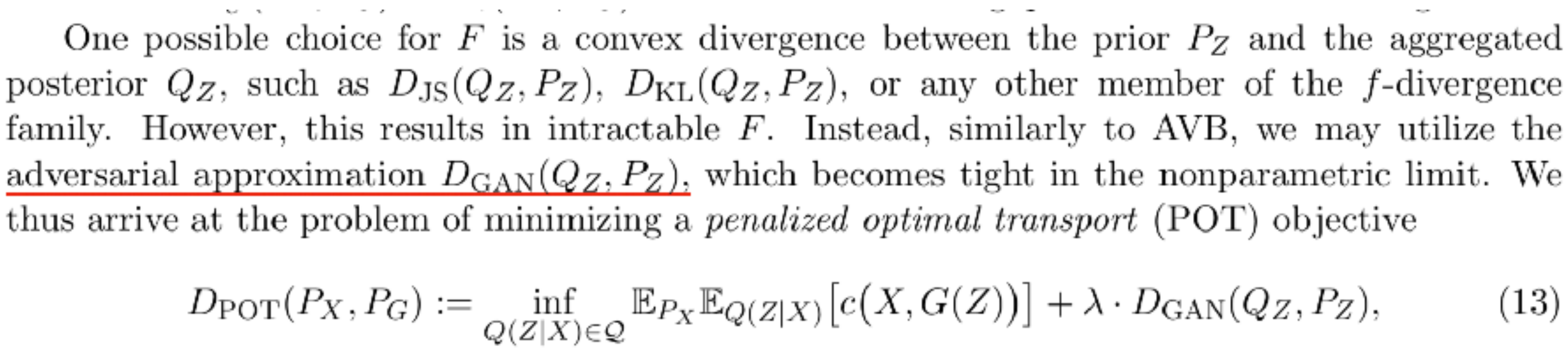


Namely, use any convex *penalty* $F\colon Q \to \mathcal{R}_+$, such that $F(Q) = 0$ if and only if $P_Z = Q_Z$, and for any $\lambda > 0$, consider the following relaxed unconstrained version of $W_c^\dagger(P_X, P_G)$:

$$W_c^\lambda(P_X, P_G) := \inf_{Q(Z|X)} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}\big[c\big(X, G(Z)\big)\big] + \lambda F(Q) \tag{12}$$



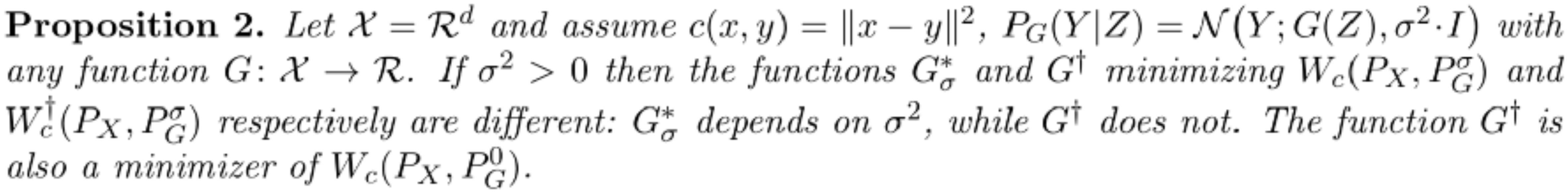
Remark 1. For $\mathcal{X} = \mathcal{R}^d$ and Gaussian $P_G(Y|Z) = \mathcal{N}(Y; G(Z), \sigma^2 \cdot I_d)$ the value of $W_c(P_X, P_G)$ is upper bounded by $W_c^\dagger(P_X, P_G)$, which coincides with the r.h.s. of (11) up to a $d \cdot \sigma^2$ additive term (see Corollary 7 in Section B.2). In other words, objective (12) coincides with the relaxed version of $W_c^\dagger(P_X, P_G)$ up to additive constant, while D_{POT} corresponds to its adversarial approximation.



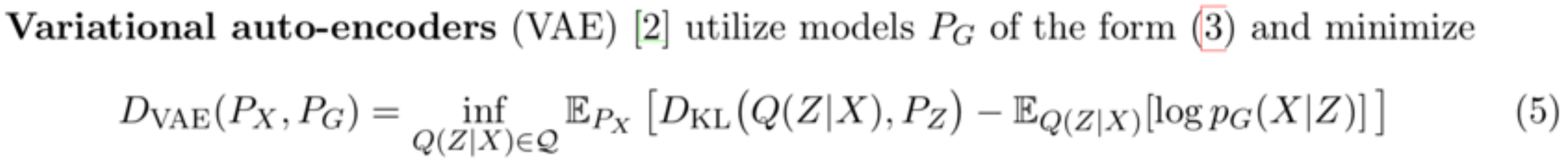
One possible choice for F is a convex divergence between the prior P_Z and the aggregated posterior Q_Z, such as $D_{\text{JS}}(Q_Z, P_Z)$, $D_{\text{KL}}(Q_Z, P_Z)$, or any other member of the f-divergence family. However, this results in intractable F. Instead, similarly to AVB, we may utilize the adversarial approximation $D_{\text{GAN}}(Q_Z, P_Z)$, which becomes tight in the nonparametric limit. We thus arrive at the problem of minimizing a *penalized optimal transport* (POT) objective

$$D_{\text{POT}}(P_X, P_G) := \inf_{Q(Z|X) \in \mathcal{Q}} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}\big[c\big(X, G(Z)\big)\big] + \lambda \cdot D_{\text{GAN}}(Q_Z, P_Z), \tag{13}$$

DPOT & AAE

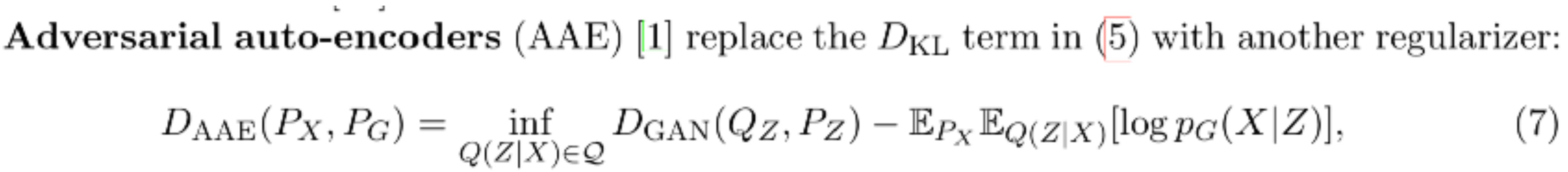


Proposition 2. *Let $\mathcal{X} = \mathcal{R}^d$ and assume $c(x,y) = \|x-y\|^2$, $P_G(Y|Z) = \mathcal{N}\big(Y; G(Z), \sigma^2 \cdot I\big)$ with any function $G\colon \mathcal{X} \to \mathcal{R}$. If $\sigma^2 > 0$ then the functions G^*_σ and $G^\dagger$ minimizing $W_c(P_X, P^\sigma_G)$ and $W^\dagger_c(P_X, P^\sigma_G)$ respectively are different: G^*_σ depends on σ^2, while $G^\dagger$ does not. The function $G^\dagger$ is also a minimizer of $W_c(P_X, P^0_G)$.*



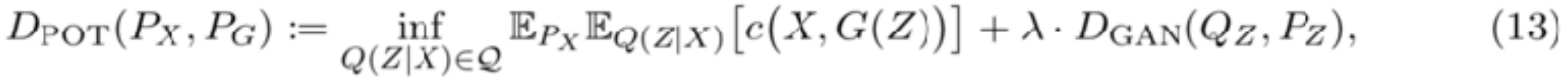
Variational auto-encoders (VAE) [2] utilize models P_G of the form (3) and minimize

$$D_{\mathrm{VAE}}(P_X, P_G) = \inf_{Q(Z|X) \in \mathcal{Q}} \mathbb{E}_{P_X}\left[D_{\mathrm{KL}}\big(Q(Z|X), P_Z\big) - \mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)]\right] \qquad (5)$$



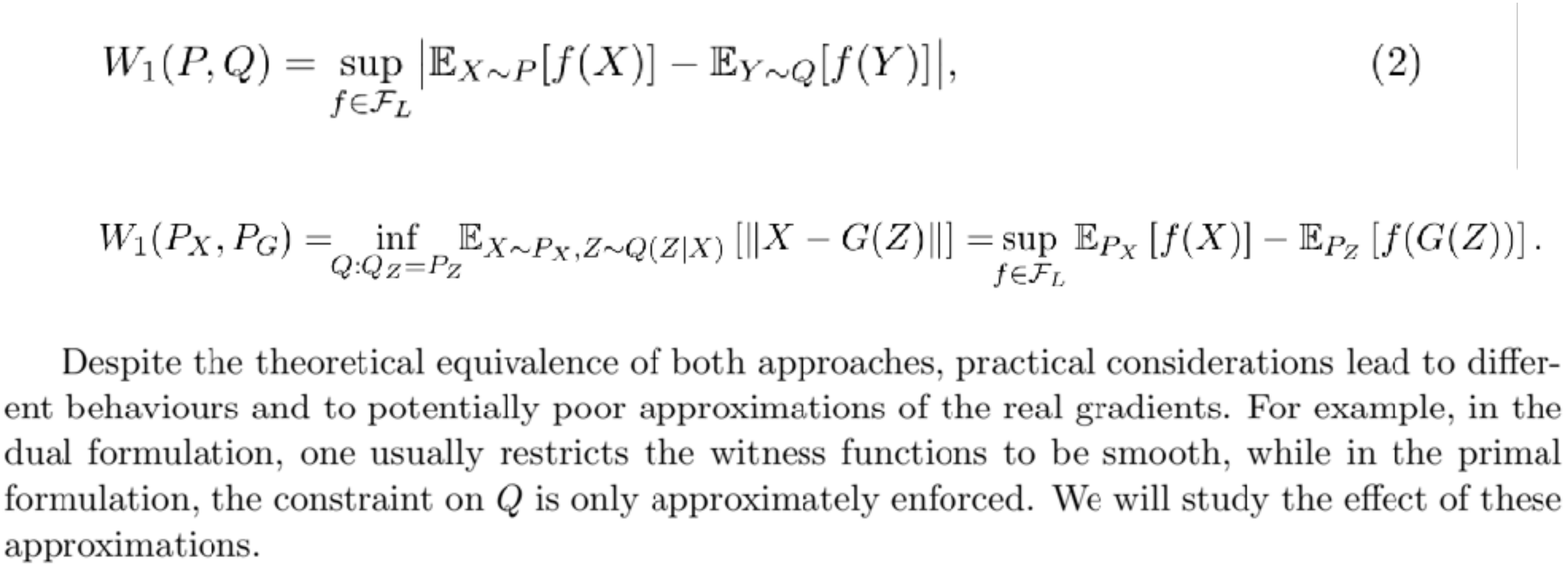
Adversarial auto-encoders (AAE) [1] replace the D_{KL} term in (5) with another regularizer:

$$D_{\mathrm{AAE}}(P_X, P_G) = \inf_{Q(Z|X) \in \mathcal{Q}} D_{\mathrm{GAN}}(Q_Z, P_Z) - \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}[\log p_G(X|Z)], \qquad (7)$$



$$D_{\mathrm{POT}}(P_X, P_G) := \inf_{Q(Z|X) \in \mathcal{Q}} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}\big[c\big(X, G(Z)\big)\big] + \lambda \cdot D_{\mathrm{GAN}}(Q_Z, P_Z), \qquad (13)$$

DPOT & WGAN



$$W_1(P, Q) = \sup_{f \in \mathcal{F}_L} \left| \mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)] \right|, \tag{2}$$

$$W_1(P_X, P_G) = \inf_{Q : Q_Z = P_Z} \mathbb{E}_{X \sim P_X, Z \sim Q(Z|X)} \left[\| X - G(Z) \| \right] = \sup_{f \in \mathcal{F}_L} \mathbb{E}_{P_X} \left[f(X) \right] - \mathbb{E}_{P_Z} \left[f(G(Z)) \right].$$

Despite the theoretical equivalence of both approaches, practical considerations lead to different behaviours and to potentially poor approximations of the real gradients. For example, in the dual formulation, one usually restricts the witness functions to be smooth, while in the primal formulation, the constraint on Q is only approximately enforced. We will study the effect of these approximations.

Thanks

Q?