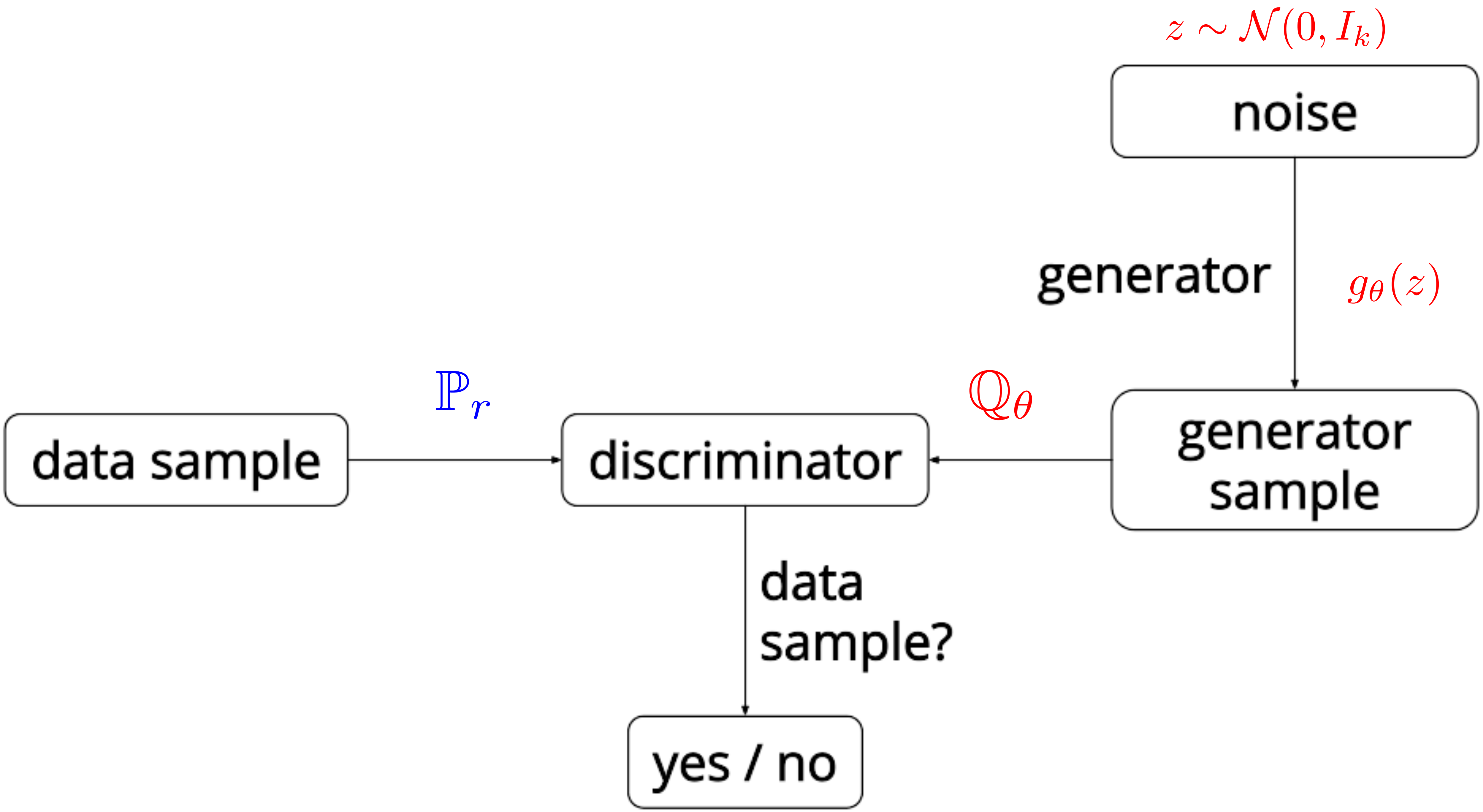
McGan: Mean and Covariance Feature Matching GAN

Youssef Mroueh*, Tom Sercu* and Vaibhava Goel

AI Foundations
IBM T.J. Watson Research Center, NY

ICML August 2017

Learning Generative Adversarial Networks



[Goodfellow et al. 2014]

Motivation: Known Issues in GAN Training

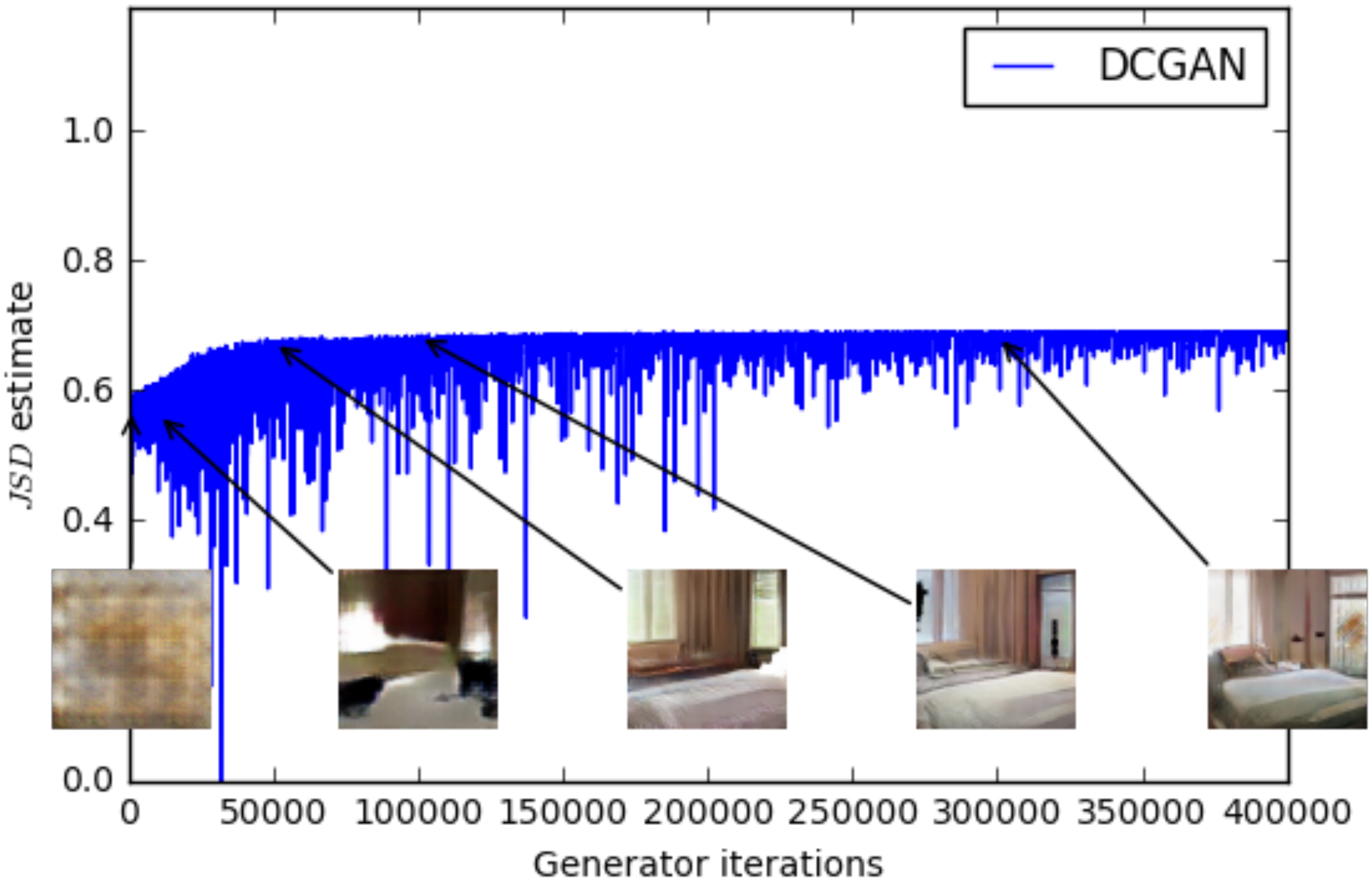


Fig from Wasserstein GAN [Arjovsky, Chintala and Bottou 2017]

Motivation: Known Issues in GAN Training

- Loss uncorrelated with sample quality.

Motivation: Known Issues in GAN Training

- Loss uncorrelated with sample quality.
- Vanishing gradient needs ad-hoc log(D(G(z))) for G training (so different loss for D and G).

Motivation: Known Issues in GAN Training

- Loss uncorrelated with sample quality.
- Vanishing gradient needs ad-hoc log(D(G(z)) for G training (so different loss for D and G).
- Optimizing a problematic metric [Arjovsky and Bottou, 2017]

Motivation: Known Issues in GAN Training

- Loss uncorrelated with sample quality.
- Vanishing gradient needs ad-hoc log(D(G(z)) for G training (so different loss for D and G).
- Optimizing a problematic metric [Arjovsky and Bottou, 2017]
- Unstable training, mode collapse.

Motivation: Known Issues in GAN Training

- Loss uncorrelated with sample quality.
- Vanishing gradient needs ad-hoc log(D(G(z)) for G training (so different loss for D and G).
- Optimizing a problematic metric [Arjovsky and Bottou, 2017]
- Unstable training, mode collapse.

We focus on GAN loss & the metric between distributions

IPMs - Integral Probability Metrics

$$d_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left\{ \mathop{\mathbb{E}}_{x \sim \mathbb{P}} f(x) - \mathop{\mathbb{E}}_{x \sim \mathbb{Q}} f(x) \right\}$$

IPMs - Integral Probability Metrics

$$d_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left\{ \mathop{\mathbb{E}}_{x \sim \mathbb{P}} f(x) - \mathop{\mathbb{E}}_{x \sim \mathbb{Q}} f(x) \right\}$$

$\mathcal{F}$ is a symmetric set of measurable, **bounded**, real valued functions.

Function class $\mathcal{F}$ defines metric.

IPMs - Integral Probability Metrics

$$d_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}} \left\{ \mathop{\mathbb{E}}_{x\sim\mathbb{P}} f(x) - \mathop{\mathbb{E}}_{x\sim\mathbb{Q}} f(x) \right\}$$

$\mathcal{F}$ is a symmetric set of measurable, **bounded**, real valued functions.

Function class $\mathcal{F}$ defines metric.

Examples:

- Maximum Mean Discrepancy:

$$\mathcal{F} = \{f \in \mathcal{H}_k \text{ RKHS }, \|f\|_{\mathcal{H}_k} \leq 1\}$$

-Wasserstein Distance: $\mathcal{F} = \{f : \|f\|_{Lip} \leq 1\}$

IPMs - Integral Probability Metrics

$$d_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}} \left\{ \mathop{\mathbb{E}}_{x\sim\mathbb{P}} f(x) - \mathop{\mathbb{E}}_{x\sim\mathbb{Q}} f(x) \right\}$$

$\mathcal{F}$ is a symmetric set of measurable, **bounded**, real valued functions.

Function class $\mathcal{F}$ defines metric.

Examples:

- Maximum Mean Discrepancy:

$$\mathcal{F} = \{f \in \mathcal{H}_k \text{ RKHS }, \|f\|_{\mathcal{H}_k} \leq 1\}$$

-Wasserstein Distance: $\mathcal{F} = \{f : \|f\|_{Lip} \leq 1\}$

Use IPM for GAN training:

$$\min_{\theta} d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{Q}_\theta)$$

Maximum Mean Discrepancy in RKHS

Φ is infinite dimensional map corresponding to kernel k

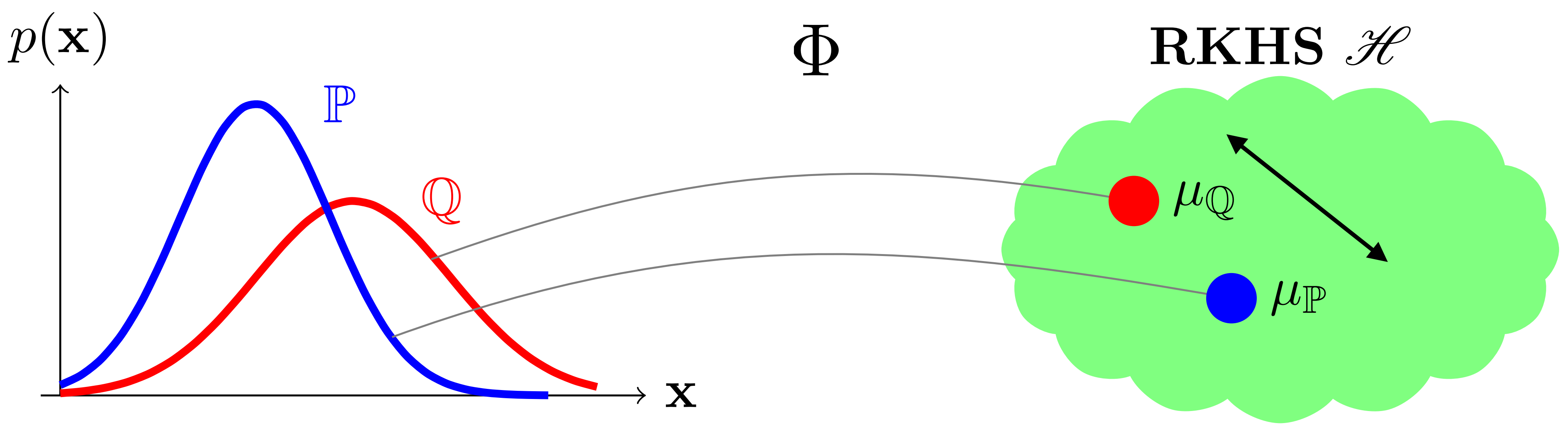
Kernel mean embedding: $\mu_{\mathbb{P}} = \mathbb{E}_{x \sim \mathbb{P}} \Phi(x)$

Maximum Mean Discrepancy in RKHS

Φ is infinite dimensional map corresponding to kernel k

Kernel mean embedding: $\mu_{\mathbb{P}} = \mathbb{E}_{x \sim \mathbb{P}} \Phi(x)$

Fig: [Muandet et al., 2016]



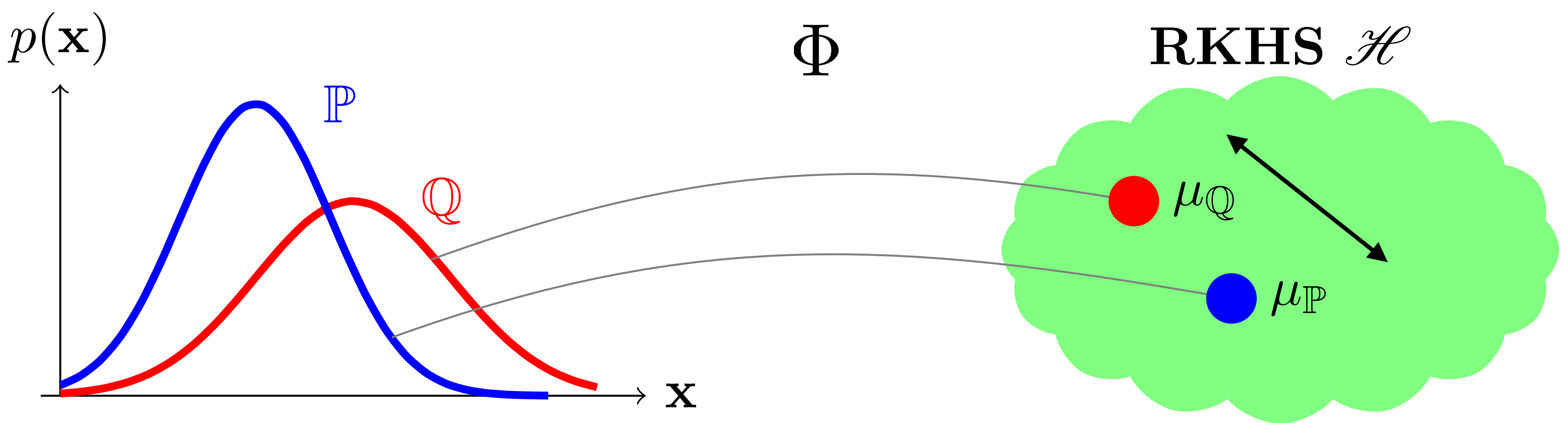
Kernel Mean Embedding of Distributions

Maximum Mean Discrepancy in RKHS

Φ is infinite dimensional map corresponding to kernel k

Kernel mean embedding: $\mu_{\mathbb{P}} = \mathbb{E}_{x\sim\mathbb{P}}\Phi(x)$

Fig: [Muandet et al., 2016]



Kernel Mean Embedding of Distributions

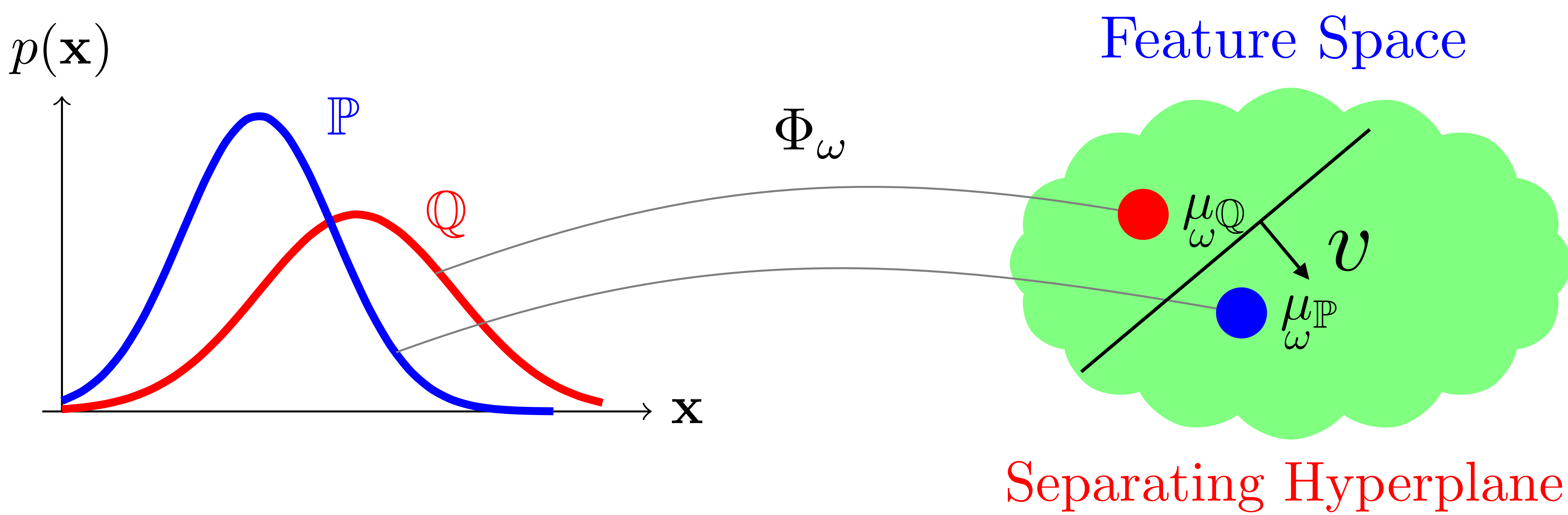
$$d_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \sup_{f, \|f\|_{\mathcal{H}}\leq 1} \langle f, \mathbb{E}_{x\sim\mathbb{P}}\Phi(x) - \mathbb{E}_{x\sim\mathbb{Q}}\Phi(x)\rangle_{\mathcal{H}}$$
$$= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$$

Mean Feature Matching IPM

McGan [Mroueh, Sercu and Goel 2017]

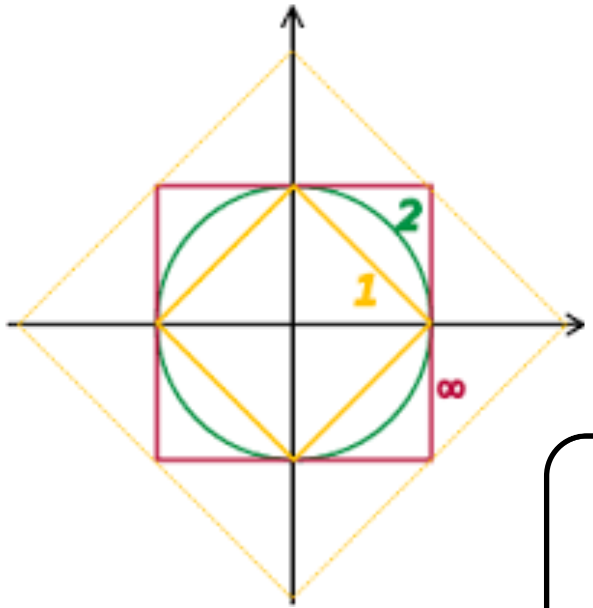
Neural Network Embedding of Distributions

Neural Network $\Phi_\omega : \mathcal{X} \to \mathbb{R}^m, \omega \in \Omega$



A Neural Network Embedding of the distribution

Refresher: dual norms



$$\max_{\|v\|_p \le 1} \langle v, z \rangle = \|z\|_q$$

$$\frac{1}{p} + \frac{1}{q} = 1$$

Mean Feature Matching IPM

$$d_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left\{ \mathop{\mathbb{E}}_{x \sim \mathbb{P}} f(x) - \mathop{\mathbb{E}}_{x \sim \mathbb{Q}} f(x) \right\}$$

Specify NN $\Phi_\omega(x)$ and $f(x)$ linear in the feature space.

$$\mathcal{F} = \{ f(x) = \langle v, \Phi_\omega(x) \rangle \}$$

Mean Feature Matching IPM

$$d_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left\{ \mathop{\mathbb{E}}_{x \sim \mathbb{P}} f(x) - \mathop{\mathbb{E}}_{x \sim \mathbb{Q}} f(x) \right\}$$

Specify NN $\Phi_\omega(x)$ and $f(x)$ linear in the feature space.

$$\mathcal{F} = \{ f(x) = \langle v, \Phi_\omega(x) \rangle \}$$

$$d_{\mathcal{F}_{v,\omega,p}}(\mathbb{P}, \mathbb{Q}) = \max_{\omega \in \Omega} \max_{v, ||v||_p \leq 1} \left\langle v, \underbrace{\mathop{\mathbb{E}}_{x \sim \mathbb{P}} \Phi_\omega(x)}_{\mu_w(\mathbb{P})} - \mathop{\mathbb{E}}_{x \sim \mathbb{Q}} \Phi_\omega(x) \right\rangle$$

Primal

Mean Feature Matching IPM

$$d_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left\{ \mathop{\mathbb{E}}_{x \sim \mathbb{P}} f(x) - \mathop{\mathbb{E}}_{x \sim \mathbb{Q}} f(x) \right\}$$

Specify NN $\Phi_\omega(x)$ and $f(x)$ linear in the feature space.

$$\mathcal{F} = \{ f(x) = \langle v, \Phi_\omega(x) \rangle \}$$

$$d_{\mathcal{F}_{v,\omega,p}}(\mathbb{P}, \mathbb{Q}) = \max_{\omega \in \Omega} \max_{v, ||v||_p \leq 1} \left\langle v, \underbrace{\mathop{\mathbb{E}}_{x \sim \mathbb{P}} \Phi_\omega(x)}_{\mu_w(\mathbb{P})} - \mathop{\mathbb{E}}_{x \sim \mathbb{Q}} \Phi_\omega(x) \right\rangle$$

Primal

$$d_{\mathcal{F}_{v,\omega,p}}(\mathbb{P}, \mathbb{Q}) = \max_{\omega \in \Omega} \| \mu_\omega(\mathbb{P}) - \mu_\omega(\mathbb{Q}) \|_q$$

$$\frac{1}{p} + \frac{1}{q} = 1$$

Dual

Mean Feature Matching IPM

$$d_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}} \left\{ \mathop{\mathbb{E}}_{x\sim\mathbb{P}} f(x) - \mathop{\mathbb{E}}_{x\sim\mathbb{Q}} f(x) \right\}$$

Specify NN $\Phi_\omega(x)$ and $f(x)$ linear in the feature space.

$$\mathcal{F} = \{f(x) = \langle v, \Phi_\omega(x)\rangle\}$$

$$d_{\mathcal{F}_{v,\omega,p}}(\mathbb{P},\mathbb{Q}) = \max_{\omega\in\Omega} \max_{v,||v||_p\leq 1} \left\langle v, \underbrace{\mathop{\mathbb{E}}_{x\sim\mathbb{P}} \Phi_\omega(x)}_{\mu_w(\mathbb{P})} - \mathop{\mathbb{E}}_{x\sim\mathbb{Q}} \Phi_\omega(x) \right\rangle$$

Primal

$$d_{\mathcal{F}_{v,\omega,p}}(\mathbb{P},\mathbb{Q}) = \max_{\omega\in\Omega} \|\mu_\omega(\mathbb{P}) - \mu_\omega(\mathbb{Q})\|_q$$

$$\frac{1}{p} + \frac{1}{q} = 1$$

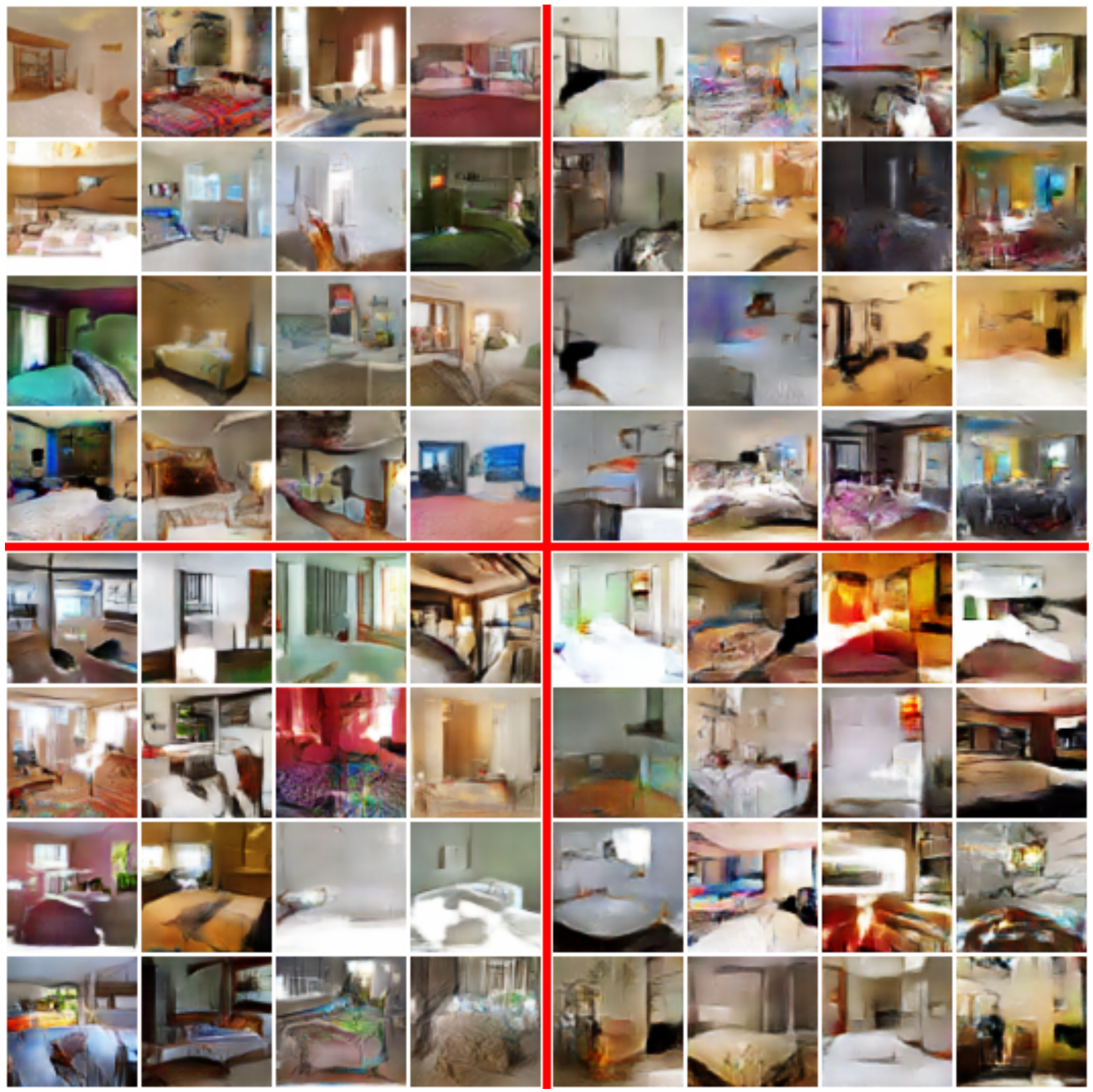
Dual

Experiments

Primal Dual

$\text{IPM}_{\mu,1}$

$\text{IPM}_{\mu,2}$

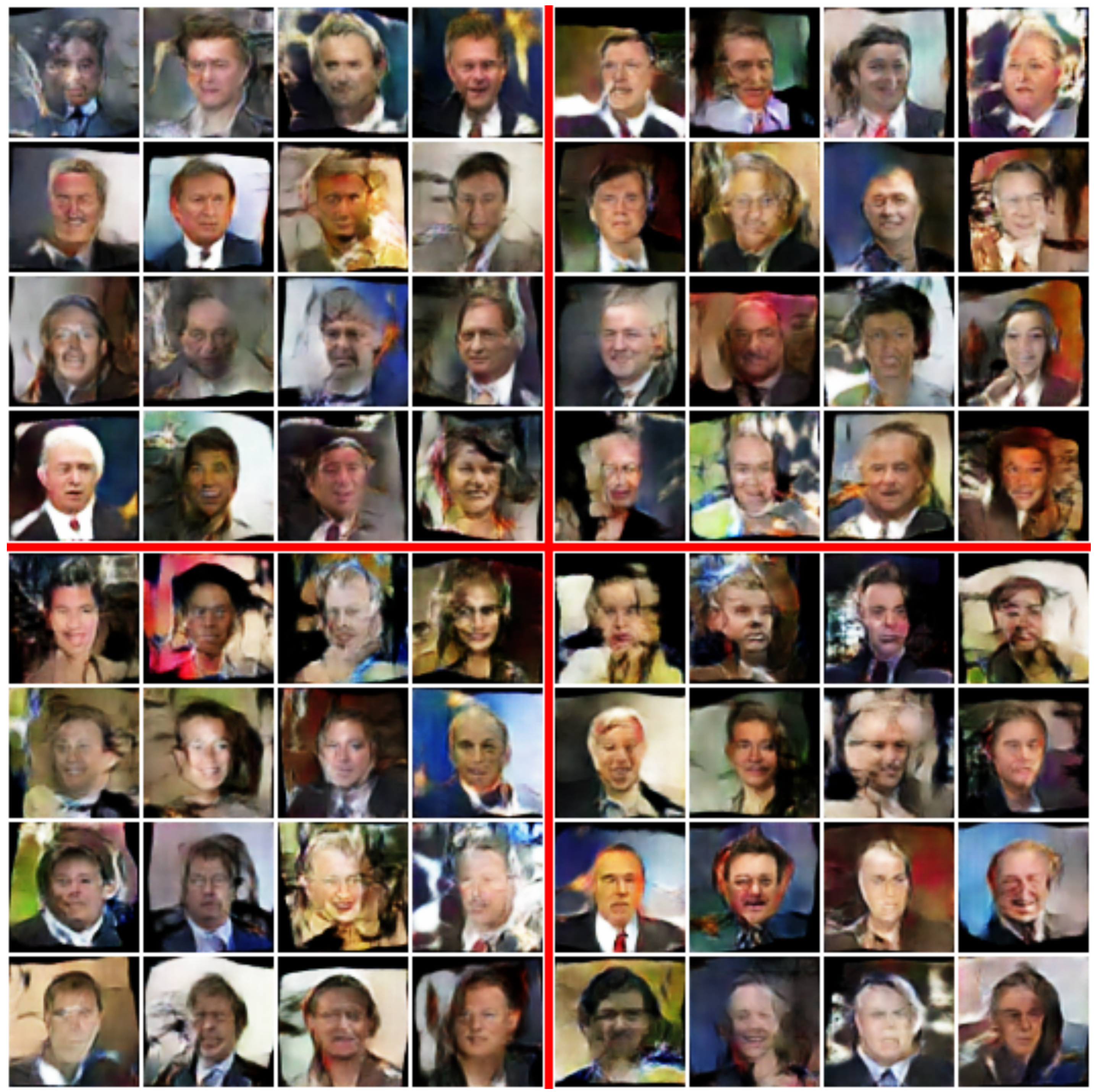


Experiments

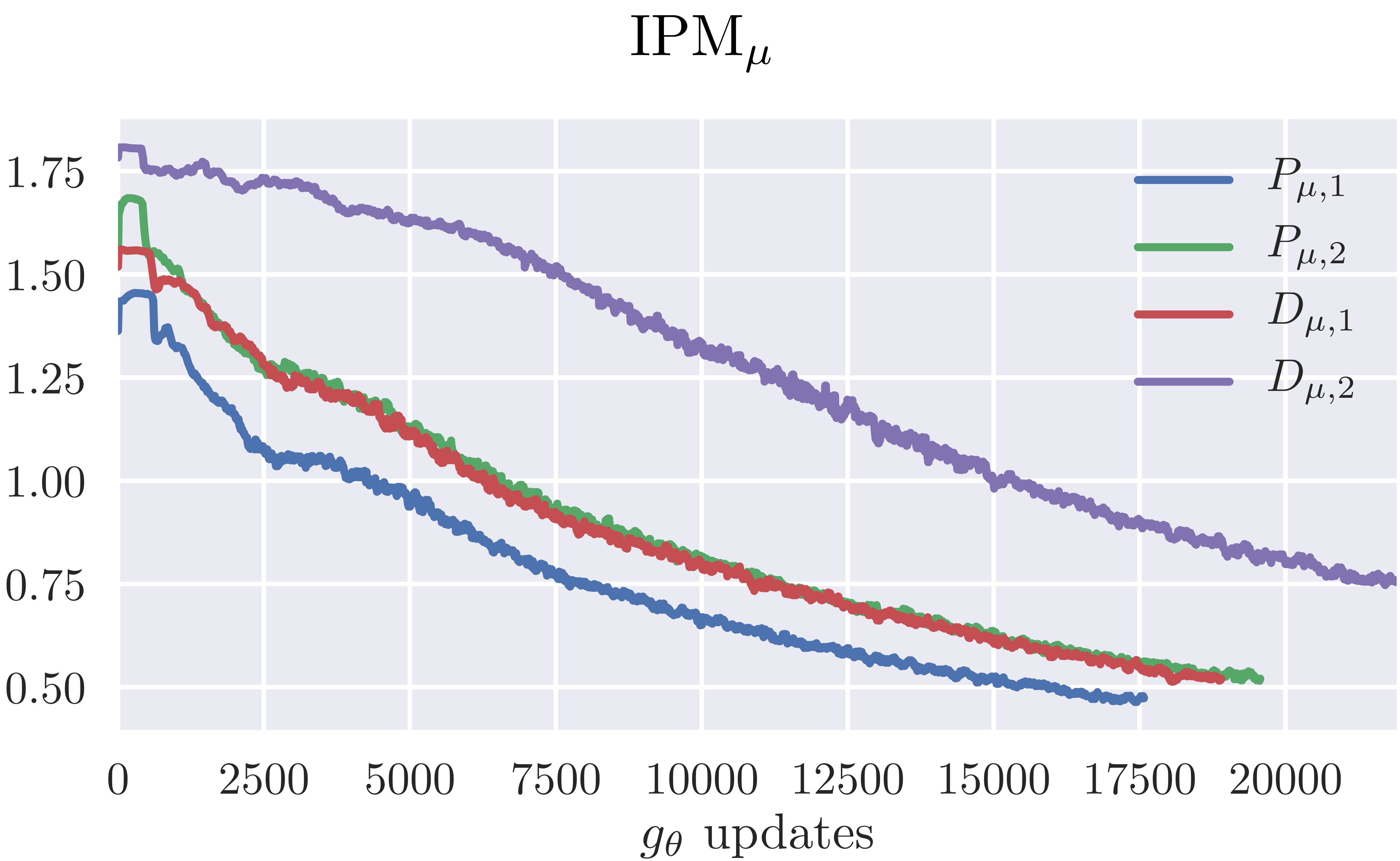
Primal Dual

$\mathrm{IPM}_{\mu,1}$

$\mathrm{IPM}_{\mu,2}$



Experiments

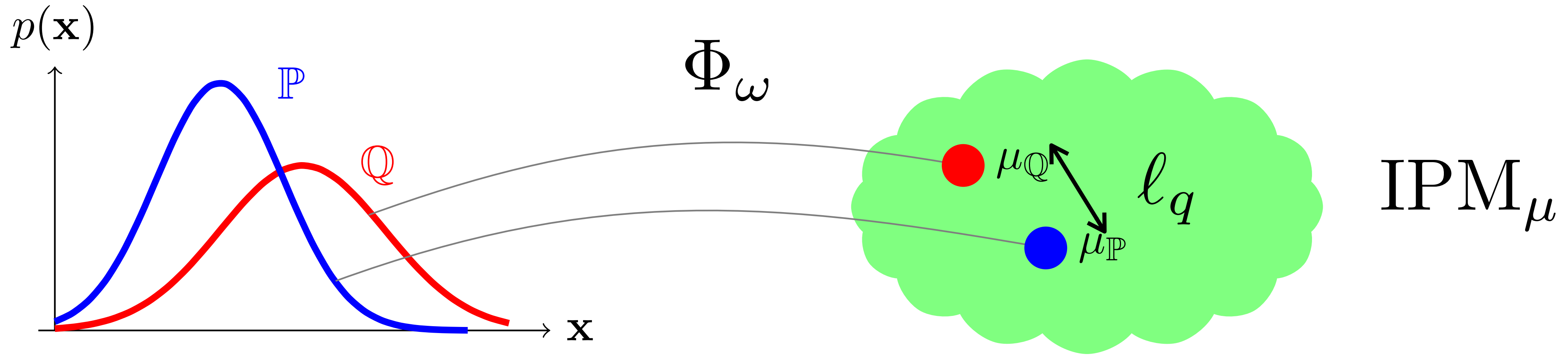


Covariance Feature Matching IPM

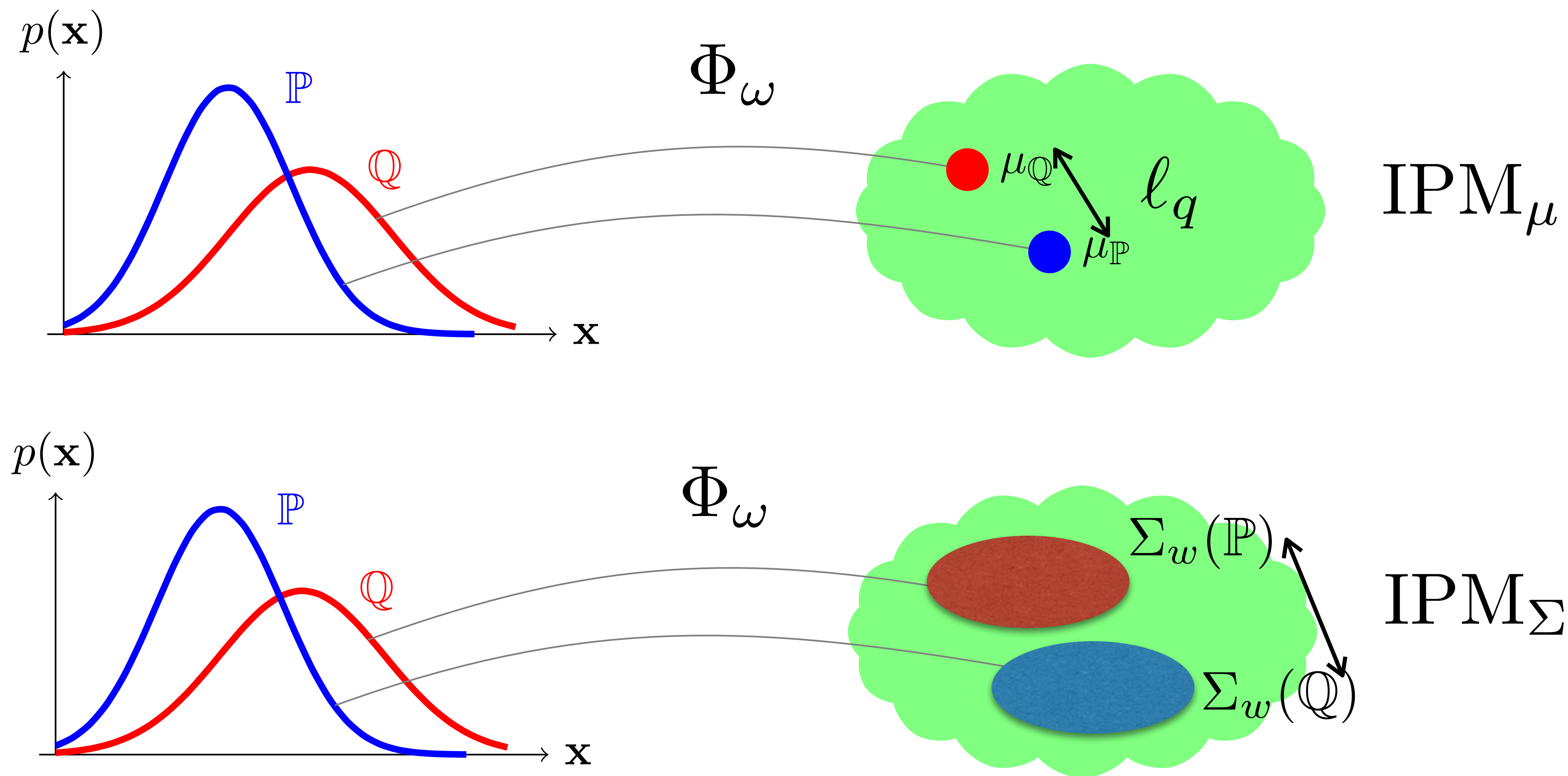
McGan [Mroueh, Sercu and Goel 2017]

Covariance Feature Matching GAN

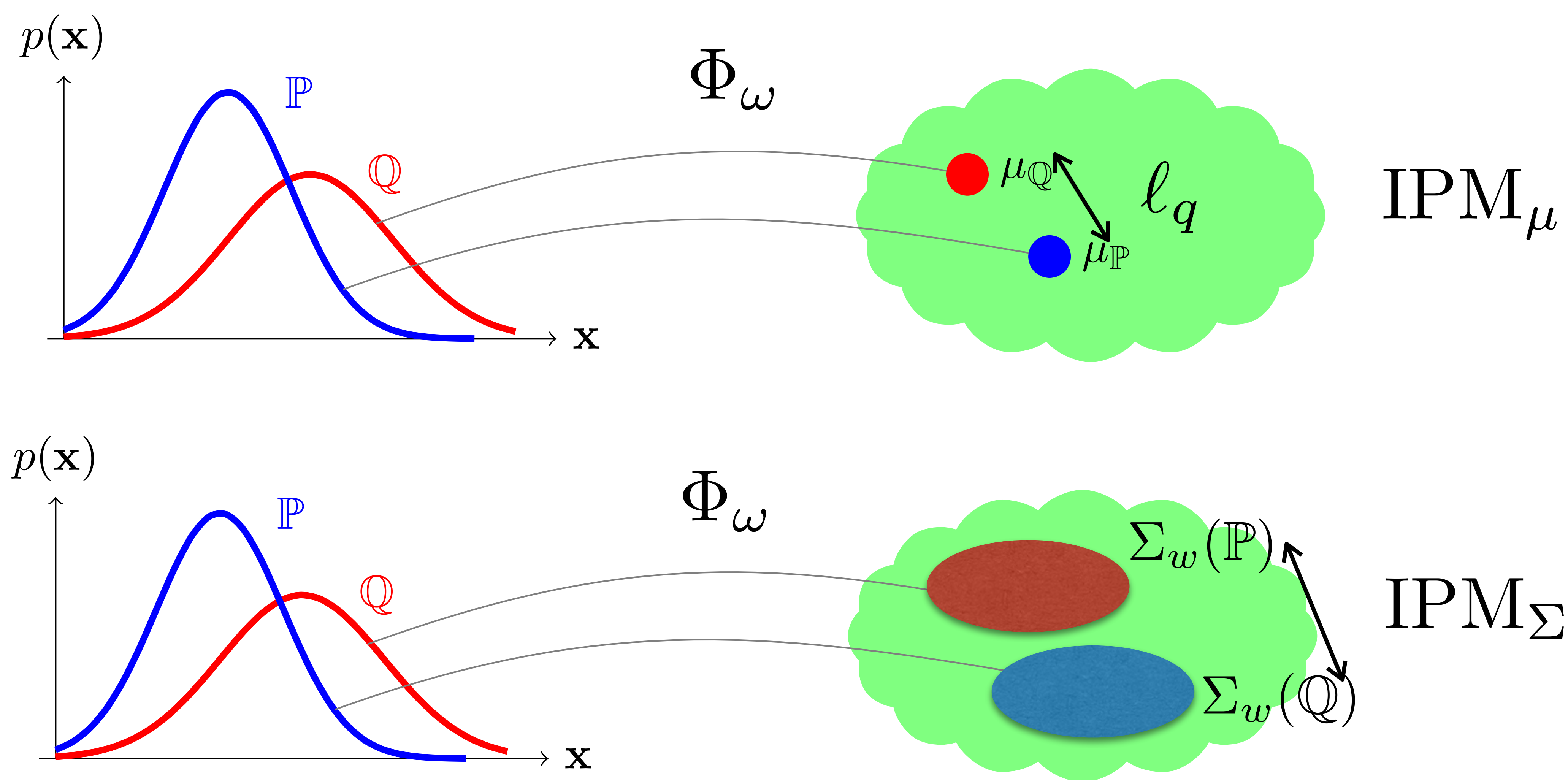
Covariance Feature Matching GAN



Covariance Feature Matching GAN

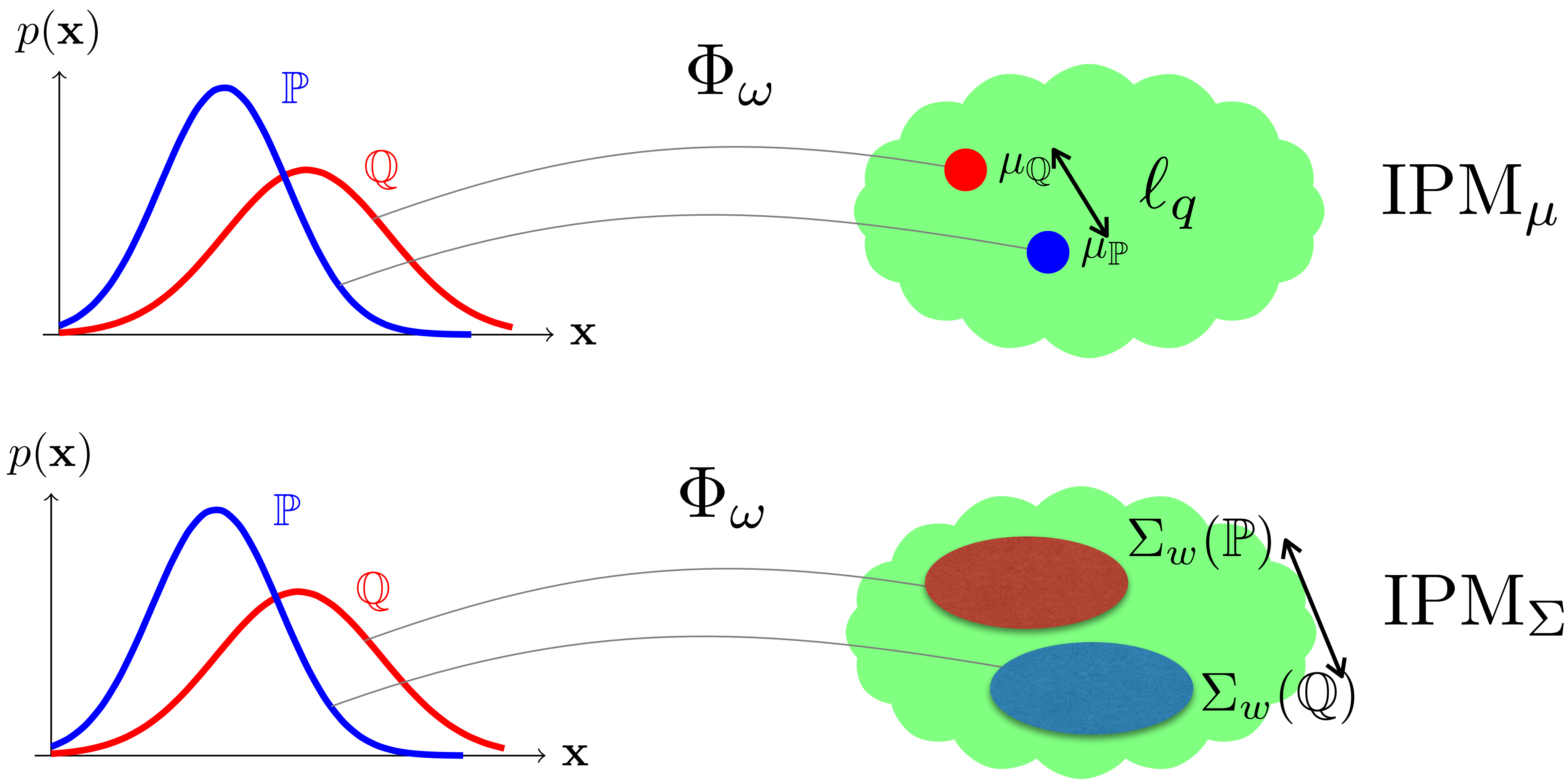


Covariance Feature Matching GAN



Next Up: Covariance matching $\|\Sigma_\omega(\mathbb{P}) - \Sigma_\omega(\mathbb{Q})\|$

Covariance Feature Matching GAN



Next Up: Covariance matching $\|\Sigma_\omega(\mathbb{P}) - \Sigma_\omega(\mathbb{Q})\|$

As an IPM!

Covariance Feature Matching GAN

Introduce $2 \times K$ orthonormal vectors $\{u_1, \ldots, u_K\}$ and $\{v_1, \ldots, v_K\}$.

And assemble into $U = [u_1|u_2|\ldots|u_K] \quad V = [v_1|v_2|\ldots|v_K]$.

$$
\begin{aligned}
f(x) &= \left\langle U^\top \Phi_\omega(x), V^\top \Phi_\omega(x) \right\rangle \\
&= \sum_{j=1}^{K} [u_j^\top \Phi_\omega(x)][\Phi_\omega(x)^\top v_j] \\
&= \mathrm{Trace}\ (\mathrm{U}^\top [\Phi_\omega(\mathrm{x})\Phi_\omega(\mathrm{x})^\top]\mathrm{V})
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{F} = \{ f_{\omega,U,V}(x) \,|\, U, V \in \mathbb{R}^{m \times K},\ & \\
U^\top U = I_K,\ & \\
V^\top V = I_K,\ & \\
\omega \in \Omega \} &
\end{aligned}
$$

Covariance Feature Matching GAN

Introduce $2 \times K$ orthonormal vectors $\{u_1, \ldots, u_K\}$ and $\{v_1, \ldots, v_K\}$.

And assemble into $U = [u_1|u_2|\ldots|u_K]$ $\quad V = [v_1|v_2|\ldots|v_K]$.

$$
\begin{aligned}
f(x) &= \left\langle U^\top \Phi_\omega(x), V^\top \Phi_\omega(x) \right\rangle \\
&= \sum_{j=1}^{K} u_j^\top [\Phi_\omega(x)\Phi_\omega(x)^\top] v_j \\
&= \mathrm{Trace}\ (\mathrm{U}^\top [\Phi_\omega(\mathrm{x})\Phi_\omega(\mathrm{x})^\top]\mathrm{V})
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{F} = \{ f_{\omega,U,V}(x) \,|\, & U, V \in \mathbb{R}^{m \times K}, \\
& U^\top U = I_K, \\
& V^\top V = I_K, \\
& \omega \in \Omega \}
\end{aligned}
$$

Covariance Feature Matching GAN

Introduce $2 \times K$ orthonormal vectors $\{u_1, \ldots, u_K\}$ and $\{v_1, \ldots, v_K\}$.

And assemble into $U = [u_1|u_2|\ldots|u_K]$ $\quad V = [v_1|v_2|\ldots|v_K]$.

$$\begin{aligned} f(x) &= \left\langle U^\top \Phi_\omega(x), V^\top \Phi_\omega(x) \right\rangle \\ &= \sum_{j=1}^{K} u_j^\top [\Phi_\omega(x)\Phi_\omega(x)^\top] v_j \\ &= \text{Trace}\ (\mathrm{U}^\top [\Phi_\omega(\mathrm{x})\Phi_\omega(\mathrm{x})^\top]\mathrm{V}) \end{aligned}$$

$$\begin{aligned} \mathcal{F} = \{ f_{\omega,U,V}(x) \,|\, & U, V \in \mathbb{R}^{m \times K}, \\ & U^\top U = I_K, \\ & V^\top V = I_K, \\ & \omega \in \Omega \} \end{aligned}$$

$$\begin{aligned} \mathop{\mathbb{E}}_{x\sim\mathbb{P}} f(x) &= \mathop{\mathbb{E}}_{x\sim\mathbb{P}} \text{Trace}\ (\mathrm{U}^\top [\Phi_\omega(\mathrm{x})\Phi_\omega(\mathrm{x})^\top]\mathrm{V}) \\ &= \text{Trace}\ (\mathop{\mathbb{E}}_{\mathrm{x}\sim\mathbb{P}} \mathrm{U}^\top [\Phi_\omega(\mathrm{x})\Phi_\omega(\mathrm{x})^\top]\mathrm{V}) \\ &= \text{Trace}\ (\mathrm{U}^\top \mathop{\mathbb{E}}_{\mathrm{x}\sim\mathbb{P}} \left[\Phi_\omega(\mathrm{x})\Phi_\omega(\mathrm{x})^\top\right] \mathrm{V}) \\ &= \text{Trace}\ (\mathrm{U}^\top \Sigma_\omega(\mathbb{P}) \mathrm{V}) \end{aligned}$$

Covariance Feature Matching GAN

Introduce $2 \times K$ orthonormal vectors $\{u_1, \ldots, u_K\}$ and $\{v_1, \ldots, v_K\}$.

And assemble into $U = [u_1|u_2|\ldots|u_K] \quad V = [v_1|v_2|\ldots|v_K]$.

$$
\begin{aligned}
f(x) &= \langle U^\top \Phi_\omega(x), V^\top \Phi_\omega(x) \rangle \\
&= \sum_{j=1}^{K} u_j^\top [\Phi_\omega(x)\Phi_\omega(x)^\top] v_j \\
&= \mathrm{Trace}\ (\mathrm{U}^\top [\Phi_\omega(\mathrm{x})\Phi_\omega(\mathrm{x})^\top] \mathrm{V})
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{F} = \{ f_{\omega,U,V}(x) \,|\, & U, V \in \mathbb{R}^{m \times K}, \\
& U^\top U = I_K, \\
& V^\top V = I_K, \\
& \omega \in \Omega \}
\end{aligned}
$$

$$
\begin{aligned}
\mathop{\mathbb{E}}_{x \sim \mathbb{P}} f(x) &= \mathop{\mathbb{E}}_{x \sim \mathbb{P}} \mathrm{Trace}\ (\mathrm{U}^\top [\Phi_\omega(\mathrm{x})\Phi_\omega(\mathrm{x})^\top] \mathrm{V}) \\
&= \mathrm{Trace}\ (\mathop{\mathbb{E}}_{\mathrm{x} \sim \mathbb{P}} \mathrm{U}^\top [\Phi_\omega(\mathrm{x})\Phi_\omega(\mathrm{x})^\top] \mathrm{V}) \\
&= \mathrm{Trace}\ (\mathrm{U}^\top \boxed{\mathop{\mathbb{E}}_{\mathrm{x} \sim \mathbb{P}} \left[\Phi_\omega(\mathrm{x})\Phi_\omega(\mathrm{x})^\top\right]} \ \mathrm{V}) \\
&= \mathrm{Trace}\ (\mathrm{U}^\top \boxed{\Sigma_\omega(\mathbb{P})} \mathrm{V})
\end{aligned}
$$

Covariance Feature Matching GAN

$$d_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \left\{ \mathop{\mathbb{E}}_{x \sim \mathbb{P}} f(x) - \mathop{\mathbb{E}}_{x \sim \mathbb{Q}} f(x) \right\}$$

$$f(x) = \text{Trace}\ (\mathrm{U}^{\top}[\Phi_{\omega}(\mathrm{x})\Phi_{\omega}(\mathrm{x})^{\top}]\mathrm{V})$$

Covariance Feature Matching GAN

$$d_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}} \left\{ \mathop{\mathbb{E}}_{x\sim\mathbb{P}} f(x) - \mathop{\mathbb{E}}_{x\sim\mathbb{Q}} f(x) \right\}$$

$$f(x) = \mathrm{Trace}\ (\mathrm{U}^{\top}[\Phi_{\omega}(\mathrm{x})\Phi_{\omega}(\mathrm{x})^{\top}]\mathrm{V})$$

$$d_{\mathcal{F}_{U,V,\omega}}(\mathbb{P},\mathbb{Q}) = \max_{\omega\in\Omega} \quad \max_{U,V\in\mathcal{O}_{m,k}} \quad Trace\left[U^{\top}(\Sigma_{\omega}(\mathbb{P}) - \Sigma_{\omega}(\mathbb{Q}))V\right]$$

Primal

Covariance Feature Matching GAN

$$d_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}} \left\{ \mathop{\mathbb{E}}_{x\sim\mathbb{P}} f(x) - \mathop{\mathbb{E}}_{x\sim\mathbb{Q}} f(x) \right\}$$

$$f(x) = \mathrm{Trace}\ (\mathrm{U}^\top[\Phi_\omega(\mathrm{x})\Phi_\omega(\mathrm{x})^\top]\mathrm{V})$$

$$d_{\mathcal{F}_{U,V,\omega}}(\mathbb{P},\mathbb{Q}) = \max_{\omega\in\Omega} \quad \max_{U,V\in\mathcal{O}_{m,k}} \quad Trace\left[U^\top(\Sigma_\omega(\mathbb{P}) - \Sigma_\omega(\mathbb{Q}))V\right]$$

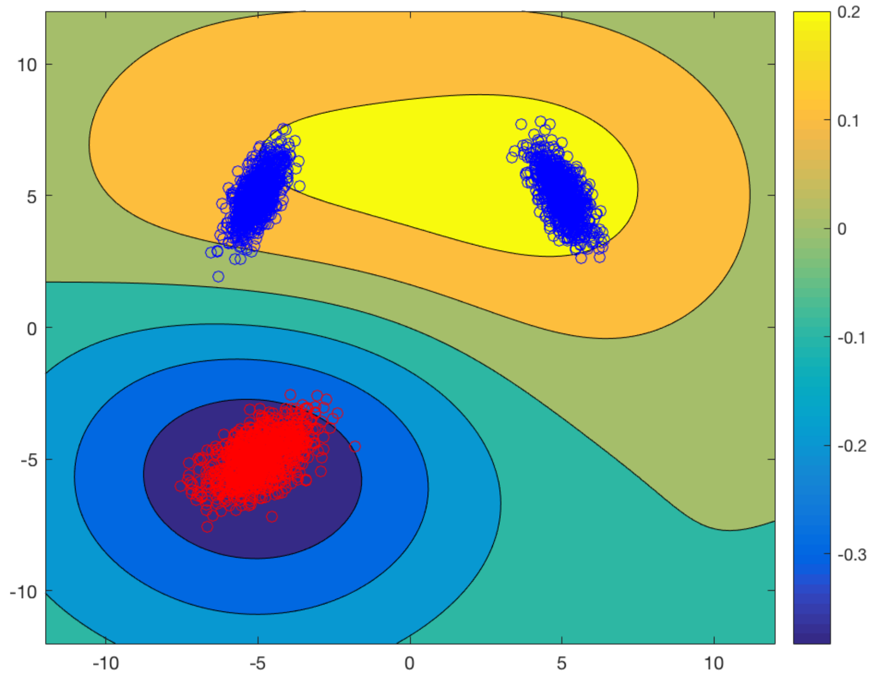
Primal

$$d_{\mathcal{F}_{U,V,\omega}}(\mathbb{P},\mathbb{Q}) = max_{\omega\in\Omega} \left\|[\Sigma_\omega(\mathbb{P}) - \Sigma_\omega(\mathbb{Q})]_k\right\|_*$$

Truncated Nuclear Norm

Dual

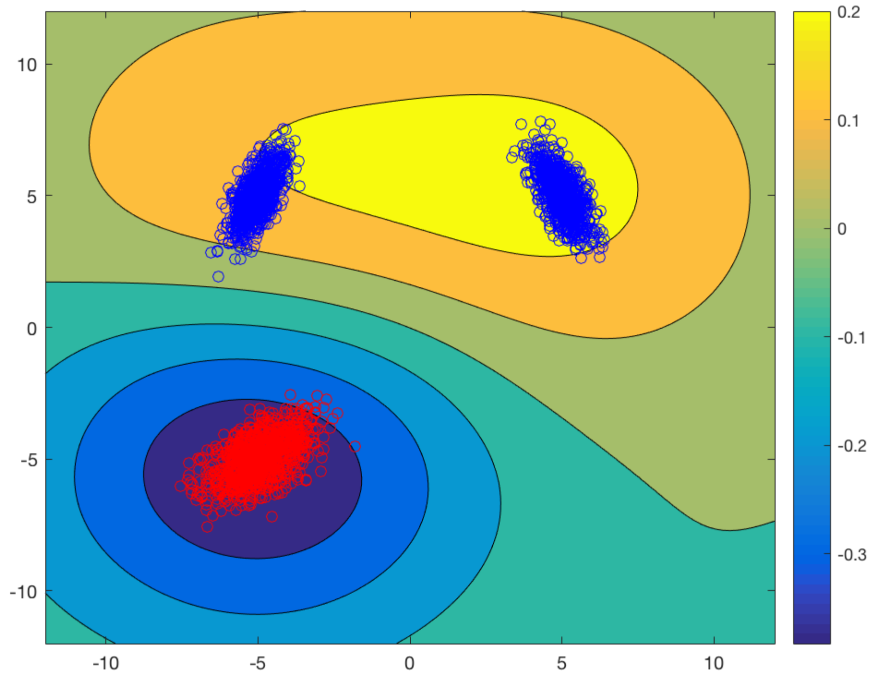
Covariance Feature Matching GAN



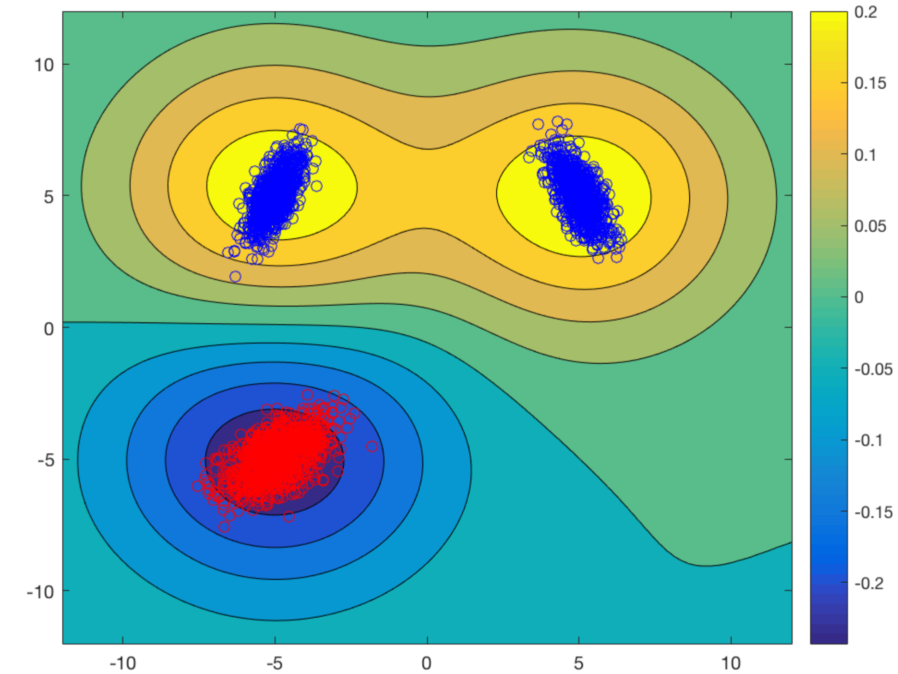
a) IPM $_{\mu,2}$: Level sets of $f(x) = \langle v^*, \Phi_\omega(x) \rangle$

$$v^* = \frac{\mu_w(\mathbb{P}) - \mu_w(\mathbb{Q})}{\|\mu_w(\mathbb{P}) - \mu_w(\mathbb{Q})\|_2}.$$

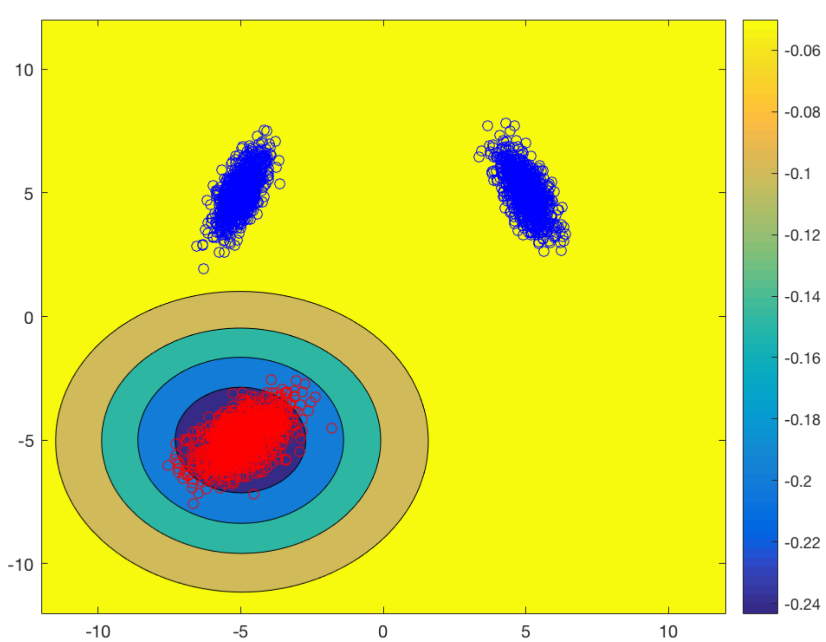
Covariance Feature Matching GAN



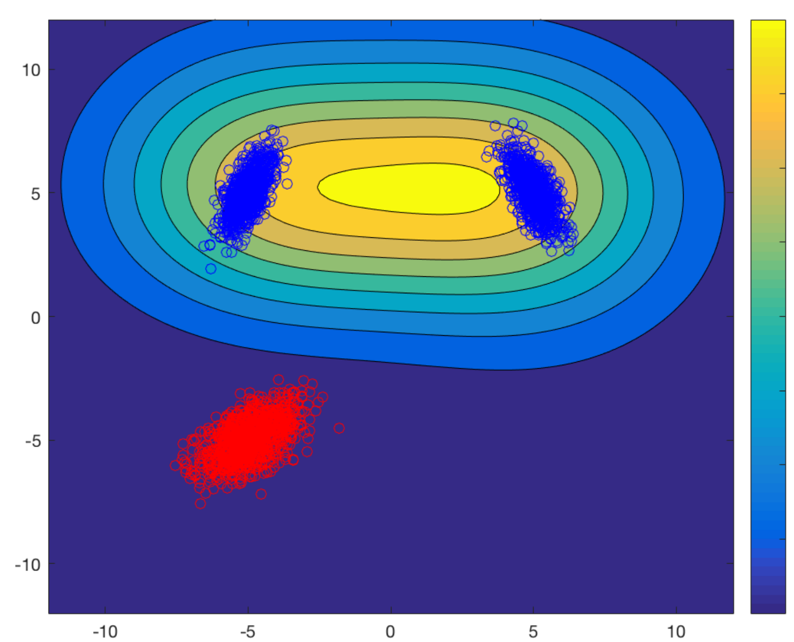
a) IPM $_{\mu,2}$: Level sets of $f(x) = \langle v^*, \Phi_\omega(x) \rangle$
$v^* = \frac{\mu_w(\mathbb{P}) - \mu_w(\mathbb{Q})}{\|\mu_w(\mathbb{P}) - \mu_w(\mathbb{Q})\|_2}$.



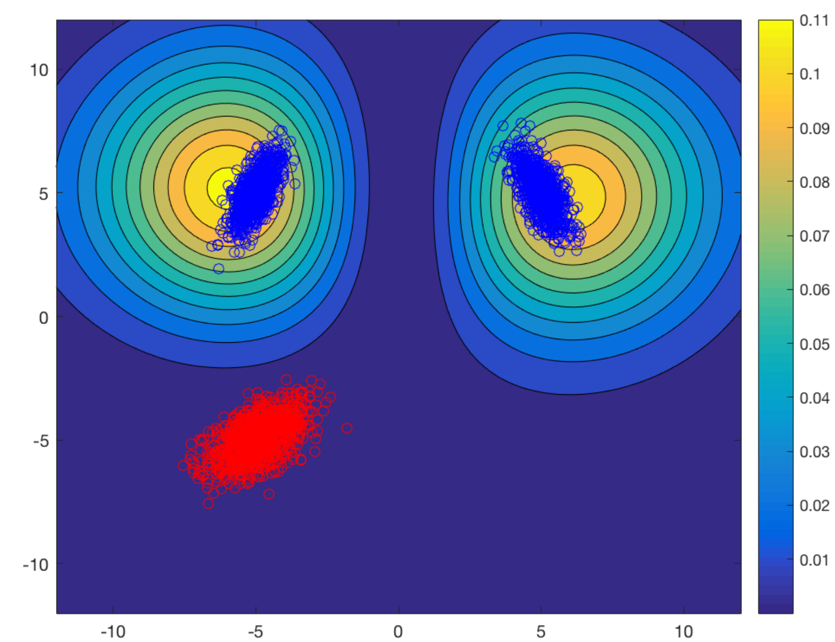
b) IPM $_\Sigma$: Level sets of $f(x) = \sum_{j=1}^{k} \langle u_j, \Phi_\omega(x) \rangle \langle v_j, \Phi_\omega(x) \rangle$
$k = 3, u_j, v_j$ left and right singular vectors of $\Sigma_w(\mathbb{P}) - \Sigma_w(\mathbb{Q})$.



Level Sets of c) $\langle u_1, \Phi_\omega(x) \rangle \langle v_1, \Phi_\omega(x) \rangle$



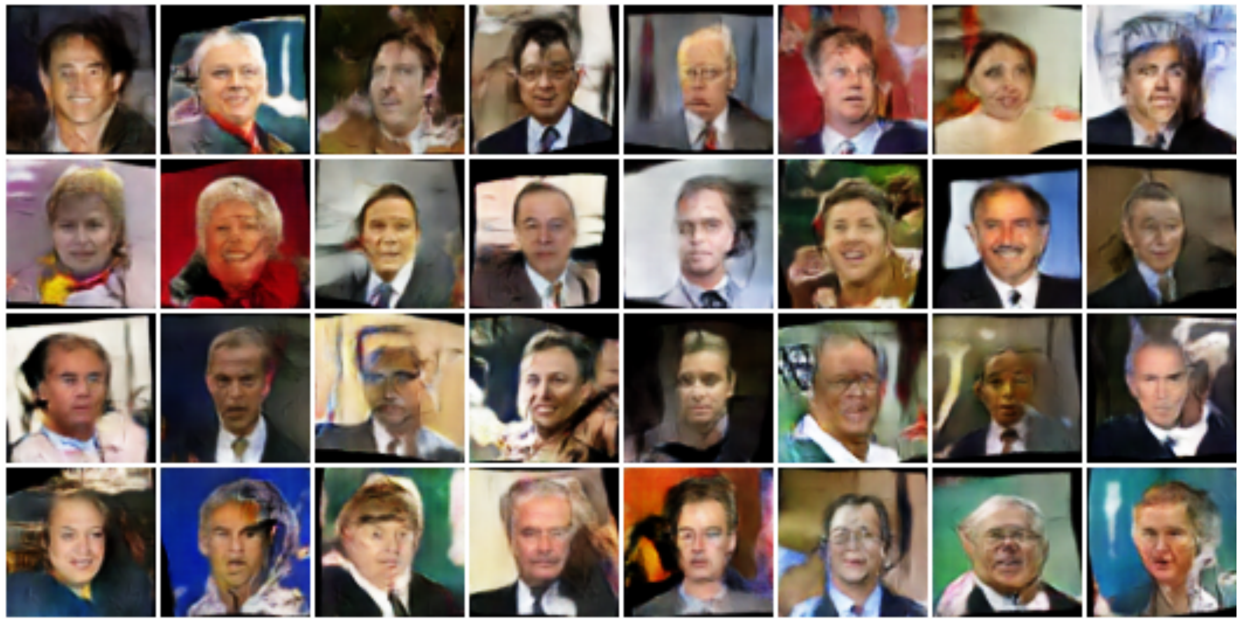
d) $\langle u_2, \Phi_\omega(x) \rangle \langle v_2, \Phi_\omega(x) \rangle$

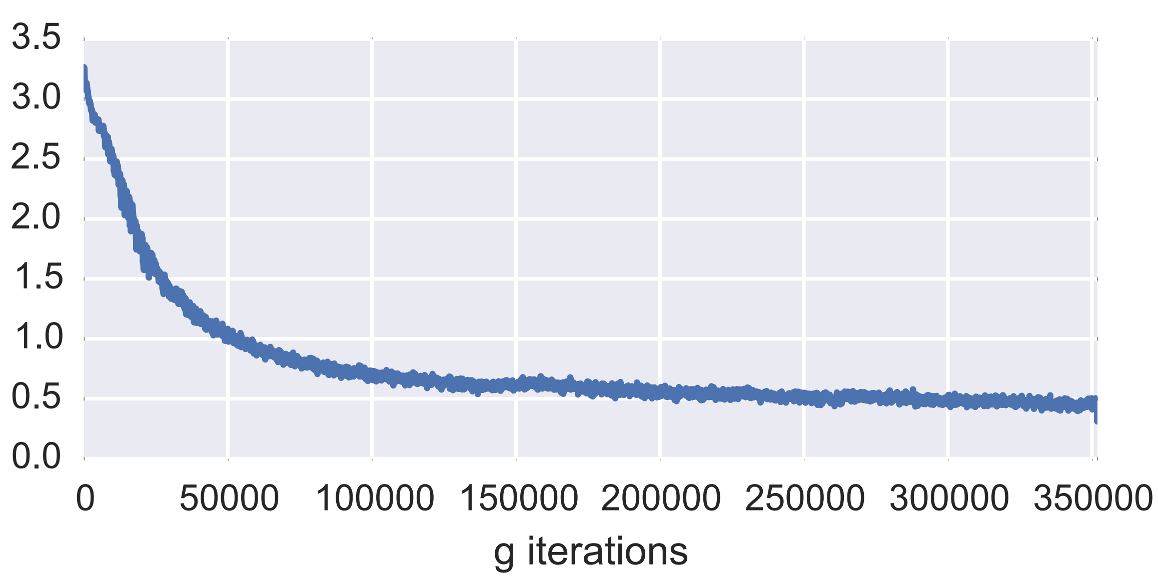


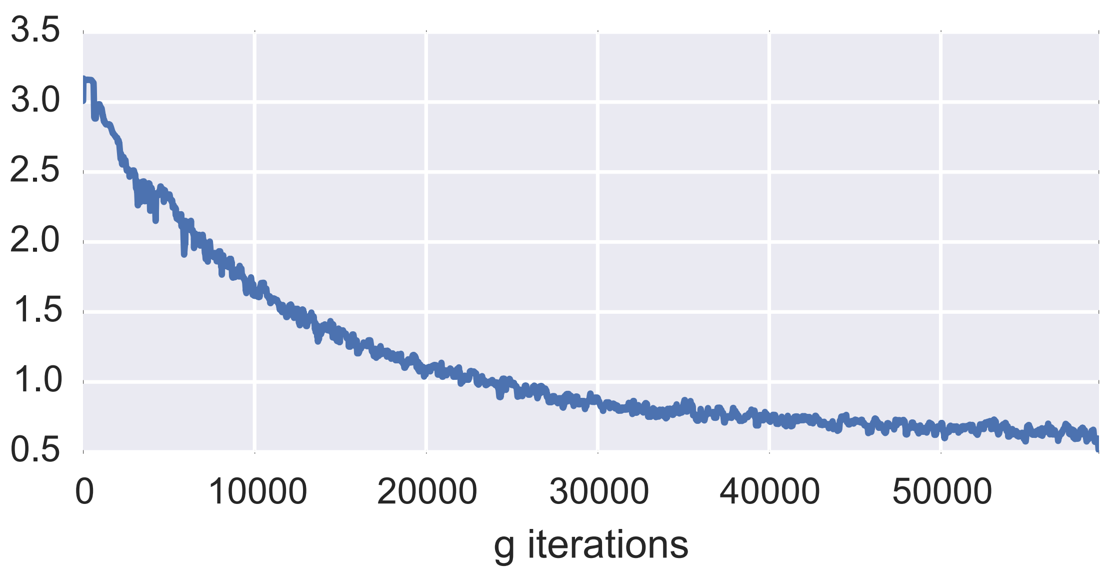
e) $\langle u_3, \Phi_\omega(x) \rangle \langle v_3, \Phi_\omega(x) \rangle$

Experiments

IPM_{Σ}



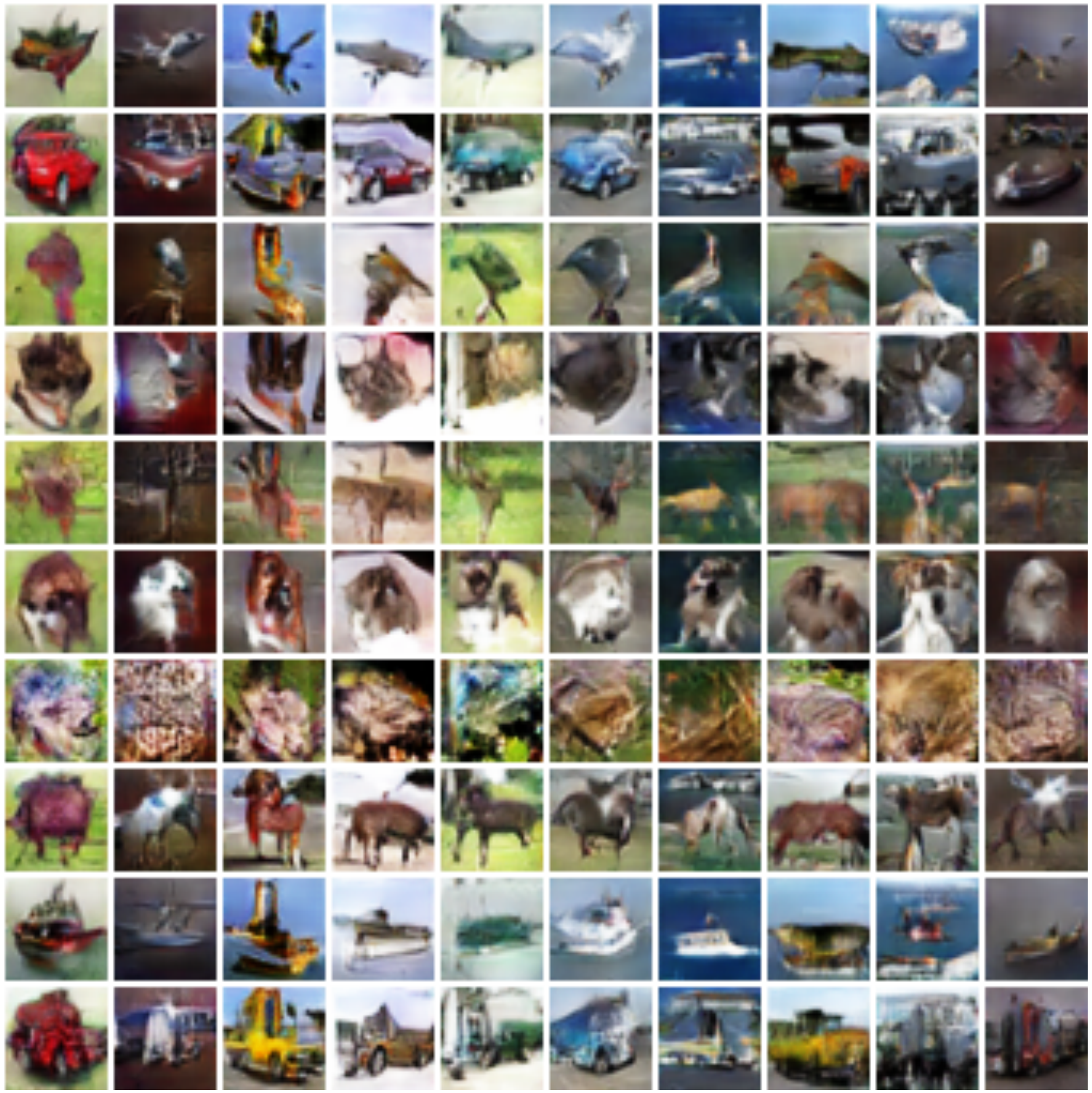




Experiments

Conditional generation using the labels on Cifar , with an auxiliary classifier (CE term)

$$\mathrm{IPM}_{\Sigma}\ ,\ k = 16$$



Experiments

Cifar-10 Inception scores of our models and baselines.

	Cond (+L)	Uncond (+L)	Uncond (-L)
L1+Sigma	7.11 ± 0.04	6.93 ± 0.07	6.42 ± 0.09
L2+Sigma	7.27 ± 0.04	6.69 ± 0.08	6.35 ± 0.04
Sigma	$\mathbf{7.29 \pm 0.06}$	$\mathbf{6.97 \pm 0.10}$	$\mathbf{6.73 \pm 0.04}$
WGAN	3.24 ± 0.02	5.21 ± 0.07	6.39 ± 0.07
BEGAN [Berthelot et al., 2017]			5.62
Impr. GAN "-LS" [Salimans et al., 2016]		6.83 ± 0.06	
Impr. GAN Best [Salimans et al., 2016]		8.09 ± 0.07	

DCGAN architecture, 32x32, with 3 extra layers.

Covariance Feature Matching GAN

Covariance Feature Matching GAN

- Efficient: no need to compute the covariance matrix explicitly!

Covariance Feature Matching GAN

- Efficient: no need to compute the covariance matrix explicitly!
- k-components as in PCA, targeting modes of the data.

Covariance Feature Matching GAN

- Efficient: no need to compute the covariance matrix explicitly!
- k-components as in PCA, targeting modes of the data.
- Meaningful and stable loss between distributions.

Questions?