# 2017 Formatting Instructions for Authors Using LaTeX

**AAAI Press**
Association for the Advancement of Artificial Intelligence
2275 East Bayshore Road, Suite 160
Palo Alto, California 94303

**Abstract**

## The Proposed $S$ Distance

We consider the one dimensional case as a start, where $x_r$ are real samples sampled from distribution $\mathbb{P}_r$, and $x_g$ are generated samples sampled from distribution $\mathbb{P}_g$,

$$x_r \sim \mathbb{P}_r \tag{1}$$

$$x_g \sim \mathbb{P}_g \tag{2}$$

Note that both $x_r$ and $x_g$ are restricted between $[0, 1]$. Following is the proposed $S$ distance,

$$S(\mathbb{P}_r, \mathbb{P}_g) = \mathbb{E}_{x_g \sim \mathbb{P}_g}\{|\int_{x_g}^{1} \mathbb{P}_r(x)dx - \int_{x_g}^{1} \mathbb{P}_g(x)dx|\} \tag{3}$$

while the Wasserstein distance is defined to be,

$$W(\mathbb{P}_r, \mathbb{P}_g) = \sup_{\|f\|_L \leq 1} \{\mathbb{E}_{x_r \sim \mathbb{P}_r}[f(x_r)] - \mathbb{E}_{x_g \sim \mathbb{P}_g}[f(x_g)]\} \tag{4}$$

Apparently, both $S$ and $W$ distance will be minimized if the $\mathbb{P}_r$ and $\mathbb{P}_g$ are identical. To take a deeper insight of the advantage of the proposed $S$ distance, we consider the representation of these two distance at one specific sample $x_g$. This is crucial, since when updating *Generator* $G$, it only observe at a specific $x_g$, instead of having a whole sight of the distributions $\mathbb{P}_r$ and $\mathbb{P}_g$. The $S$ at $x_g$ is,

$$S_{\mathbb{P}_r, \mathbb{P}_g}(x_g) = |\int_{x_g}^{1} \mathbb{P}_r(x)dx - \int_{x_g}^{1} \mathbb{P}_g(x)dx| \tag{5}$$

while the $W$ at $x_g$ is,

$$W_{\mathbb{P}_r, \mathbb{P}_g}(x_g) = f(x) \approx \mathbb{P}_r(x) - \mathbb{P}_g(x) \tag{6}$$

We can see that $S_{\mathbb{P}_r, \mathbb{P}_g}(x_g)$ consider how unbalance are the two distributions in a whole sight, while the $W_{\mathbb{P}_r, \mathbb{P}_g}(x_g)$ considers the unbalance of the two probabilities at this specific $x_g$. One can easily think of a $x_g$, where $W_{\mathbb{P}_r, \mathbb{P}_g}(x_g)$ is zero, but the distributions $\mathbb{P}_r$ and $\mathbb{P}_r$ are not identical, which means $W_{\mathbb{P}_r, \mathbb{P}_g}(x_g)$ is failing. But under this situation, $S_{\mathbb{P}_r, \mathbb{P}_g}(x_g)$ still gives a right direction for $x_g$ to update, since it observes the unbalance of $\mathbb{P}_r$ and $\mathbb{P}_r$ on each side of the $x_g$.



## GAN Based on $S$ Distance

Following we proposed the method to achieve this $S$ distance in a GAN. We still describe things in one-dimensional case to make it simple and straight forward. For every $x_r, x_g$ pair, we sample $x_\tau$ between $x_r$ and $x_g$,

$$x_\tau = \tau x_r + (1 - \tau)x_g \tag{7}$$

where

$$\tau \sim U[0, 1] \tag{8}$$

Following, we consider our problem on a discrete space with interval of $\varepsilon \to 0$, we give every notation of $x$ a check mark, i.e., $\check{x}$, to mark that they are discrete value under interval $\varepsilon$. Later on, we will derive limitation on $\varepsilon \to 0$, so that we can have a general conclusion on the continuous space. Now consider a event denoted by: $\check{x}_\tau \overset{t}{=} \check{x}_n$, which means,

- Sample $\check{x}_\tau$ for $t$ times, $\check{x}_n$ got sampled as $\check{x}_\tau$ at least for one time.

To be clear, $\check{x}_\tau$ $\check{x}_r$, $\check{x}_g$ are all random variables, and $\check{x}_n$ is a specific point. Apparently, we have,

$$P(\check{x}_\tau \overset{1}{=} \check{x}_n | \check{x}_r, \check{x}_g) = \begin{cases} \frac{1}{d/\varepsilon} & \check{x}_r < \check{x}_n < \check{x}_g, \check{x}_g < \check{x}_n < \check{x}_r \\ 0 & \text{else} \end{cases} \tag{9}$$

where

$$d = |\check{x}_r - \check{x}_g| \tag{10}$$

If we sample $\check{x}_\tau$ for $t$ times, where

$$t = d/\delta \tag{11}$$

Here, $\delta$ is also approaching to zero. Then, we,

$$\begin{aligned} & P(\check{x}_\tau \overset{t}{=} \check{x}_n | \check{x}_r, \check{x}_g) \\ = \; & 1 - (1 - P(\check{x}_\tau \overset{1}{=} \check{x}_n | \check{x}_r, \check{x}_g))^t \\ = \; & \begin{cases} 1 - (1 - \frac{1}{d/\varepsilon})^{d/\delta} & \check{x}_r < \check{x}_n < \check{x}_g, \check{x}_g < \check{x}_n < \check{x}_r \\ 0 & \text{else} \end{cases} \end{aligned} \tag{12}$$

We assume $\delta$ approaches zero in the same order as $\varepsilon$ approaching zero[1]. Now, we consider following limit,

$$\begin{aligned} & \lim_{\varepsilon,\delta\to 0}(1-\frac{1}{d/\varepsilon})^{d/\delta} \\ = & \lim_{\varepsilon,\delta\to 0} e^{d/\delta \ln(1-\frac{1}{d/\varepsilon})} \\ = & \lim_{\varepsilon,\delta\to 0} e^{\frac{\ln(\frac{d-\varepsilon}{d})}{\delta/d}} \\ = & \lim_{\varepsilon,\delta\to 0} e^{\frac{-1}{\frac{d-\varepsilon}{1/d}}} \\ = & e^{-1} \end{aligned} \tag{13}$$

Put the conclusion of (13) into (12), we have,

$$\begin{aligned} & P(x_\tau = x_n|x_r, x_g) = \lim_{\varepsilon,\delta\to 0} P(\check{x}_\tau \overset{t}{=} \check{x}_n|\check{x}_r, \check{x}_g) \\ = & \begin{cases} 1-e^{-1} & x_r < x_n < x_g, x_g < x_n < x_r \\ 0 & \text{else} \end{cases} \end{aligned} \tag{14}$$

where we have switch back to the continuous space and have this general conclusion. Now, we propose our update rules for the *Discriminator* $D$ with parameter $\theta$ to be optimized,

$$\theta \longrightarrow \theta + \nabla_\theta\{-|\nabla_{x_\tau} D^\theta(x_\tau) - \frac{x_r - x_g}{|x_r - x_g|}|^2\} \tag{15}$$

which means we try to make $\nabla_{x_\tau} D^\theta(x_\tau)$ approach $\frac{x_r - x_g}{|x_r - x_g|}$. Lets take a look at $\nabla_{x_\tau} D^\theta(x_\tau)$ at a specific point $x_n$,

$$\begin{aligned} & \nabla_{x_\tau = x_n} D^\theta(x_\tau = x_n) \\ = & P(x_\tau = x_n|x_g < x_n < x_r)P(x_g < x_n < x_r) \\ & -P(x_\tau = x_n|x_r < x_n < x_g)P(x_r < x_n < x_g) \end{aligned} \tag{16}$$

which means the value of $\nabla_{x_\tau = x_n} D^\theta(x_\tau = x_n)$ is determined by the probability of it gets positive update and negative update. Since (14), we know that

$$P(x_\tau = x_n|x_g < x_n < x_r) = 1 - e^{-1} \tag{17}$$

$$P(x_\tau = x_n|x_r < x_n < x_g) = 1 - e^{-1} \tag{18}$$

Put (17) (18) into (16), we have,

$$\begin{aligned} & \nabla_{x_\tau = x_n} D^\theta(x_\tau = x_n) \\ = & [P(x_g < x_n < x_r) - P(x_r < x_n < x_g)](1 - e^{-1}) \\ = & [\int_0^{x_n} \mathbb{P}_g(x)dx \int_{x_n}^1 \mathbb{P}_r(x)dx) \\ & - \int_0^{x_n} \mathbb{P}_r(x)dx \int_{x_n}^1 \mathbb{P}_g(x)dx)](1 - e^{-1}) \\ = & [\int_{x_n}^1 \mathbb{P}_r(x)dx - \int_{x_n}^1 \mathbb{P}_g(x)dx](1 - e^{-1}) \end{aligned} \tag{19}$$

Now, we can give the update rule of *Generator* $G$ with parameter $\beta$ to be learnt, and we put (19) in this update rule to have a clearer view on what $G$ is doing,

$$\begin{aligned} & \beta \\ \longrightarrow & \beta + \nabla_\beta\{D^\theta(G^\beta(x_g))\} \\ \longrightarrow & \beta + \{[\int_{x_g}^1 \mathbb{P}_r(x)dx - \int_{x_g}^1 \mathbb{P}_g(x)dx](1 - e^{-1})(\nabla_\beta G^\beta(x_g))\} \end{aligned} \tag{20}$$

which means wherever $x_g$ is, it is updating itself to make $\int_{x_g}^1 \mathbb{P}_g(x)dx$ approaching $\int_{x_g}^1 \mathbb{P}_r(x)dx$. One can just take a few cases to confirm this. The absolute error when updating $G$ is actually modelling the proposed $S$ distance in (3) (5).

[1]Note that in practice, $\varepsilon$ may approach zero in a much more higher order than $\delta$ approaching zero, i.e., $\varepsilon = a\delta^b$. Since $\varepsilon$ is the minimal data value we can have on a computer, while $\delta$ is a hyper-parameter we set to be as large as possible. But this does not effect the conclusion we will have in (13).