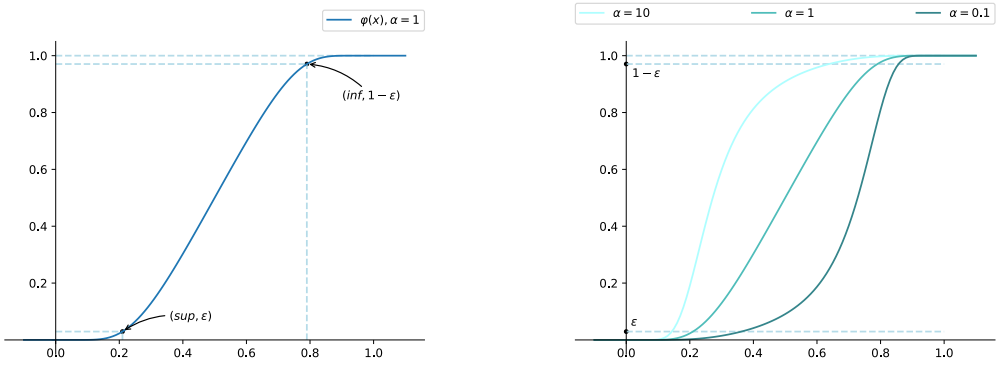
Supplementary Material for 'Balanced Spatial-Temporal Graph Structure Learning for Multivariate Time Series Forecasting: A Trade-off between Efficiency and Flexibility'

1. Description of SSU module



(a) The smooth function $\varphi(x), \alpha = 1$. (b) The curves of different α values.

Figure 1: Basic curves of SSU.

1.1. Sparsification coefficient

As we defined in context, let

$$\begin{aligned} f(x) &= \begin{cases} e^{-\frac{1}{x}} & (x > 0), \\ 0 & (x \leq 0), \end{cases} \\ \varphi(x) &= \frac{\alpha f(x)}{\alpha f(x) + f(1-x)} \ (\alpha \in \mathbb{R}_+), \end{aligned} \tag{1}$$

where parameter α is the sparsification coefficient. It can determine the shape of the curve φ and the sparse degree of the generated adjacency matrices. The sparsification effect of SSU is described below.

It is obvious that $\varphi(x) \equiv 0$, $\varphi'(x) \equiv 0$ for $x \leq 0$; $\varphi(x) \equiv 1$, $\varphi'(x) \equiv 0$ for $x \geq 1$. So we just consider $0 < x < 1$, and let $t = \dfrac{f(1-x)}{f(x)} = e^{\frac{1}{x} - \frac{1}{1-x}} := g(x)$. For $g'(x) = e^{\frac{1}{x} - \frac{1}{1-x}} \left[-\dfrac{1}{x^2} - \dfrac{1}{(1-x)^2} \right] < 0$, $g(x)$ decreases strictly monotonically on $(0, 1)$. Thus g is a

bijection and has an inverse function g^{-1}.
For $\varphi(x) < \varepsilon$, i.e.$\dfrac{\alpha}{\alpha + t} < \varepsilon$,

$$t > \alpha\left(\frac{1}{\varepsilon} - 1\right) \iff x < g^{-1}\left(\alpha\left(\frac{1}{\varepsilon} - 1\right)\right) \triangleq \sup.$$

For $\varphi(x) > 1 - \varepsilon$, i.e.$\dfrac{\alpha}{\alpha + t} > 1 - \varepsilon$,

$$t < \alpha\left(\frac{1}{1-\varepsilon} - 1\right) \iff x > g^{-1}\left(\alpha\left(\frac{1}{1-\varepsilon} - 1\right)\right) \triangleq \inf.$$

In Figure 1(b)subfigure, fixing ε, as α decreases, sup, inf increase and the length of interval $\varphi^{-1}((0, \varepsilon)) = (0, \sup)$ increase, and vice versa. If we consider the elements in the adjacency matrix $A = (A_{ij})_{n\times n}$ have a uniform distribution in $[0, 1]$, then as α decreases, the probability of a_{ij} falling into $(0, \sup)$ and A being sparse increases. Therefore, we get the conclusion that α can control the sparsification effect of SSU.

1.2. Gradient redefinition

In our experiments, as x approaches 0 and 1, the gradient approaches 0 rapidly, which leads to the vanishing gradient problem. In fact, it is extensively difficult to train the adjacency matrix values to zero and achieve the sparsification effect. Therefore, we redefine the gradient as 1 in intervals $(0, \sup)$ and $(\inf, 1)$ to accelerate convergence making the activation value fall into $\{0, 1\}$ or $(\varphi(\sup), \varphi(\inf))$ faster.