Derivation of Kernel of Dynamic Support Vector Machines:

Stochastic and Deterministic Data Process

Masamichi Sato*

ABSTRACT

We give an analytic derivation of kernel of dynamic support vector machine (DSVM). We show them for the cases of stochastic and deterministic changes. For the stochastic case, Gaussian kernel is naturally derived. For the deterministic case, the kernel is derived in the form of traveling wave. We also give comments from physical viewpoints in the context of information geometry.

*mmsato11@yahoo.co.jp

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1624v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2016, publ: 5 Jan 2016

1 Introduction

After the invention of support vector machine (SVM) [1], this method has been widely spread and used in many areas. One of the extension is the application to dynamic data. The classification of dynamic data is an important to solve actual problems. Dynamic SVM (DSVM) has been proposed in a simple extension of basic SVM [2, 3]. As further extension of SVM, support vector regression was proposed and it is used with practical purpose, such as demand forecasting [4].

In this paper, we extend the SVM framework to the classification problems of dynamic data with the other nature: though their changes are time-dependent, but they behave stochastically or deterministically. Our extension derives kernels in analytic forms within a criterion of Bellman principle. For stochastic case, Gaussian kernel is naturally derived. For deterministic case, the kernel takes the form of traveling wave.

This paper is organized as follows. In section two, we review basic SVM algorithms. In section three, we give algorithmic introduction of existing works of dynamic SVM. Section four is a main part of current paper. We give the derivations of kernel of DSVM for stochastic and deterministic cases. Section five is devoted for theoretical comments. We give comments mainly with physical viewpoints, there. Section six is conclusions.

2 Support Vector Machine

2.1 Basic SVM

A reference for this part is [5]. Suppose that the input space $\in \mathbb{R}^n$ and a set of data $\mathbf{x_1}, \ldots, \mathbf{x_r}$ are given. The identifier function is given as,

$$f(x) = w^T x - b. \tag{1}$$

The coefficient w is a weight and b is a non-negative bias. $d-1$ dimensional hypersurface identifier that satisfies $f(x) = 0$ is described as

$$\{x \in \ : \ (w^T x) + b = 0\}. \tag{2}$$

When training data $(x_1, y_1), \ldots, (x_l, y_l)$, $x_i \in , y_i \in \{\pm 1\}$, $i = 1, \ldots, l$ are given and consider the problem to identify the identifier function,

$$f_{w,b} = \text{sgn}\ ((w \cdot x) + b) \tag{3}$$

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1624v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2016, publ: 5 Jan 2016

that satisfies

$$f_{w,b}(x_i) = y_i,\ i = 1, \ldots, l. \tag{4}$$

For this identifier function, we set the constraints of following equation,

$$y_i(w^T x_i + b) \leq 1,\ i = 1, \ldots, l. \tag{5}$$

In this case, the distance between the identifier surface (margin) and these hyper surface is $\frac{1}{||w||}$. So the problem to find the parameters that maximize margin, results in finding the parameters that minimize the objective function:

$$\tau(w) = \frac{1}{2}||w||^2 \tag{6}$$

under the constraint of eq. (5). We introduce the Lagrange multipliers $\alpha_i (\geq 0)$ and rewrite the objective function as

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{l} \alpha_i \{y_i((w^T x_i) + b) - 1\}. \tag{7}$$

From partial differentiation with respect to w and b, we obtain.

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i, \tag{8}$$

$$0 = \sum_{i=1}^{l} \alpha_i y_i. \tag{9}$$

Substituting these to the objective function gives the dual problem:

$$L_D(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j x_i^T x_j, \tag{10}$$

under the constraints,

$$\sum_{i=1}^{l} \alpha_i y_i = 0. \tag{11}$$

$$0 \leq \alpha_i,\ i = 1, \ldots, l. \tag{12}$$

w is obtained from optimal α by using eq. (8) and b is given as

$$b = -\frac{1}{2}(w^T x_{+1} + w^T x_{-1}), \tag{13}$$

here, x_{+1}, x_{-1} are support vectors belonging to the class $1, -1$.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1624v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2016, publ: 5 Jan 2016

To solve the non-linear problems, we introduce non-linear mapping function. Using the non-linear function $\phi(x)$, if the inner product between x_1 and x_2 is represented as

$$\phi(x_1)^T\phi(x_2) = K(x_1, x_2), \tag{14}$$

we obtain the optimal non-linear mapping, instead of calculating characteristics $\phi(x_1)$, $\phi(x_2)$. We call such K as kernel. The frequently used kernels are the following:

polynomial kernel

$$K(x_1, x_2) = (1 + x_1^T x_2)^p \tag{15}$$

Gaussian kernel

$$K(x_1, x_2) = \exp\left(\frac{-||x_1 - x_2||^2}{2\sigma^2}\right) \tag{16}$$

Sigmoid kernel

$$K(x_1, x_2) = \tanh(ax_1^T x_2 - b) \tag{17}$$

Using kernel, the objective function is represented as

$$\begin{aligned} L_D(\alpha) = & \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i,j=1}^N \alpha_i\alpha_j t_i t_j \phi(x_i^T)\phi(x_j) \\ = & \sum_{i=1}^N \alpha_i - \frac{1}{2}\sum_{i,j=1}^N \alpha_i\alpha_j t_i t_j K(x_i, x_j), \end{aligned} \tag{18}$$

and the optimal identifier function is obtained as

$$\begin{aligned} f(x) &= \text{sign}\left(\sum_{i=1}^l \alpha_i y_i \phi(x_i)^T\phi(x) + b\right) \\ &= \text{sign}\left(\sum_{i=1}^l \alpha_i y_i K(x_i, x) + b\right) \end{aligned} \tag{19}$$

3 Dynamic SVM

3.1 Dynamic SVM Criterion

In dynamic SVM criterion [2], they assume that data and parameters are dynamic, and they consider that these transit with conditional probabilities. They assume a priori parametric distribution of the instances

$$\phi(x|a, b, y; c) = \text{const.},\ yz(a, x) \leq 1, \tag{20}$$

$$\exp(-c(1 - yz(a, x))),\ yz(a, x) < 1. \tag{21}$$

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1624v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2016, publ: 5 Jan 2016

The objective function is

$$J(a, b, \delta_1, \ldots, \delta_N | c) = a^T a + c \sum_{j=1}^{N} \delta_j \to \min, \tag{22}$$

subject to

$$y_j(a^T a + b) \leq 1 - \delta j,\ \delta j \leq 0,\ j = 1, \ldots, N. \tag{23}$$

The basic SVM criterion:

$$\begin{aligned} \phi(x|a_t, b_t, y; c) &= \text{const.},\ yz(a_t, x) \geq 1, & (24)\\ & \exp(-c(1 - yz(a_t, x))),\ yz(a_t, x) < 1, & (25) \end{aligned}$$

becomes following form for dynamic data,

$$\begin{aligned} J(a_t, b_t, \delta_{t,j},\ t = 0, \ldots, T) &= a_0^T a_0 + \frac{1}{d} \sum_{t=1}^{T} (a_t - q a_{t-1})^T (a_t - q a_{t-1}) \\ &\quad + \frac{1}{d'} \sum_{t=1}^{T} (b_t - b_{t-1})^2 + \sum_{t=1}^{T} \sum_{j=1}^{N_t} \delta_{j,t} \end{aligned} \tag{26}$$

$$\to \min_{[a_t, b_t]_{t=1}^{T}} y_{j,t}(a_t^T x_{j,t} + b_t) \leq 1 - \delta_{j,t},\ \delta_{j,t} \leq 0, j = 1, \ldots, N_t,\ t = 1, \ldots, T$$

where $z(x, a_t) = a_t^T x + b = 0$. This is the dynamic SVM criterion. After this prescription, they are considering the optimization by dynamic programming.

3.2 Distributing Kernel

The other approach to dynamic SVM is the assumption of distributing kernel [3]. The problem is represented as follows,

$$\max W(\alpha) = \sum_{i=1}^{n_x} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n_x} \alpha_i \alpha_j y_i y_j \exp\left\{ -\frac{|x_i - x_j|^2}{\sigma_i \sigma_j} \right\}, \tag{27}$$

subject to

$$0 \leq \alpha_i \leq C,\ \sum_{i=1}^{n_x} \alpha_i y_i = 0. \tag{28}$$

The identifier function is,

$$f(x) = \text{sgn}\left[\sum_{i=1}^{n_x} \exp\left\{ -\frac{|x_i - x_j|^2}{\sigma_i \sigma_j} \right\} + b^* \right]. \tag{29}$$

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1624v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2016, publ: 5 Jan 2016

Although this is the same form with basic SVM, but the difference is the assumption of data which follows dynamic process. The data transit from x_i to x_j with the changes of time from i to j, following the distribution of

$$\exp\left\{-\frac{|x_i - x_j|^2}{\sigma_i \sigma_j}\right\}. \tag{30}$$

4 Stochastic and Deterministic DSVMs

Here we consider the two types processes of dynamic data: stochastic and deterministic. The former transit stochastically depending on time. The transition of latter is deterministic, but it depends on time. We derive the appropriate forms of kernel for each processes. For the stochastic case, Gaussian functional form of Kernel is naturally derived. For deterministic case, it takes the form of traveling wave.

4.1 Stochastic DSVM

4.1.1 Setting of Problem for Stochastic DSVM

We assume a stochastic behavior for data process as depicted in Fig. 1.

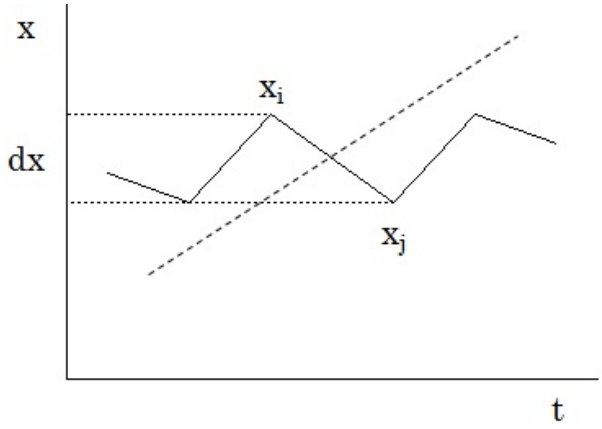

Figure 1: Stochastic DSVM

Our objective is detecting the change of tendency of stochastic process, represented as classification problem. The process changes as

$$x_j = x_i + dx, \tag{31}$$

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1624v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2016, publ: 5 Jan 2016

and the inifinitismal change follows a stochastic process,

$$dx = \mu dt + \sigma dz, \tag{32}$$

here dt is time difference and dz is standard Brownian motion. μ is growth rate and σ is volatility. The objective function takes the following form,

$$L_D(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j J(x_i, x_j), \tag{33}$$

4.1.2 Kernel of Stochastic DSVM

The optimization problem is equivalent to consider the following problem,

$$\max \left[J(x, T) \right] \tag{34}$$

Here, T is terminal time. By Bellman principle [6], solving this problem results in solving the sub-problem,

$$J(x) = \max \mathbf{E}_{(x;t)}[J(x + dx)]. \tag{35}$$

Here, $\mathbf{E}_{(x;t)}$ represents taking conditional expectation at (x, t). This equation is transformed as

$$\max \mathbf{E}_{(x;t)}[J(x + dx) - J(x)] = 0 \tag{36}$$

For inifinitismal difference, this equation becomes,

$$J(x + dx) - J(x_t) = J'(x)dx + \frac{1}{2} J''(x)dx^2 + o(dt). \tag{37}$$

By Ito's lemma,

$$\mathbf{E}_{(x;t)}[J(x + dx) - J(x)] = \mu J'(x)dx + \frac{1}{2}\sigma^2 J''(x)dt + o(dt) \tag{38}$$

This derives a diffusion equation such as

$$\frac{\partial J(x)}{\partial t} - \mu \frac{\partial}{\partial x} J(x) - \frac{1}{2}\sigma^2 \frac{\partial^2}{\partial x^2} J(x) = 0 \tag{39}$$

we obtain the solution of the following form (with the boundary condition $J(\infty) = 0$),

$$J(x) = \frac{1}{2\sqrt{\pi\sigma^2(t - \mu^2/2\sigma^2)}} \exp\left(-\frac{(x - \mu/\sigma^2)^2}{4\sigma^2(t - \mu^2/2\sigma^2)} \right). \tag{40}$$

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1624v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2016, publ: 5 Jan 2016

If $\mu = 0$, then

$$J(x) = \frac{1}{2\sqrt{\pi\sigma^2 t}} \exp(-\frac{x^2}{4\sigma^2 t}). \tag{41}$$

This is the same functional form as Gaussian kernel. So, the objective function becomes the following form:

$$L_D(\alpha(t)) = \sum_{i=1}^{l} \alpha_i(t) - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i(t)\alpha_j(t+\Delta t) y_i y_j J(x_i(t), x_j(t+\Delta t)), \tag{42}$$

here,

$$J(x_i(t), x_j(t+\Delta t)) = \frac{1}{2\sqrt{\pi\sigma^2 \Delta t}} \exp\left(-\frac{||x_i(t) - x_j(t+\Delta t)||^2}{4\sigma^2 \Delta t}\right). \tag{43}$$

As shown in above, the same functional form as Gaussian kernel was naturally derived. This gives one of the reasonings why basic SVM works. They are supposing stochastic distribution to data implicitly. Above derivation is a verification of its effectiveness.

4.2 Deterministic DSVM

4.2.1 Setting of Problem for Deterministic DSVM

The generalization to deterministic case of basic SVM is obtained by the introduction of time dependence,

$$x_j = x_i + dx, \tag{44}$$

and dx follows,

$$dx^2 = c^2 dt^2. \tag{45}$$

What we consider is the classification problem as depicted in Fig. 2. The objective function is,

$$L_D(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \tag{46}$$

4.2.2 Kernel of Deterministic DSVM

The Bellman principle for current case is described as

$$J(x) = \max J(x + dx). \tag{47}$$

This equation derives

$$\max\left[J(x + dx) - J(x)\right] \tag{48}$$

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1624v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2016, publ: 5 Jan 2016

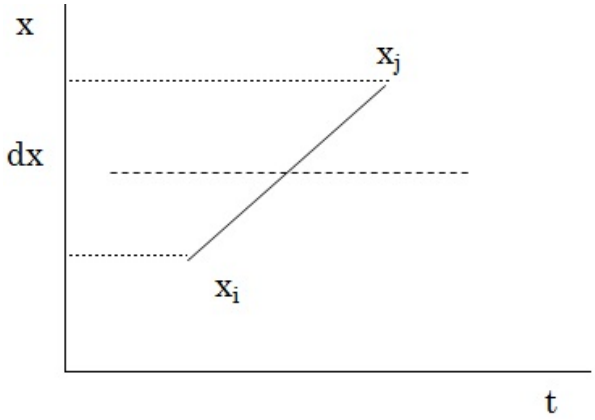

Figure 2: Deterministic DSVM

For inifinitismal difference, this equation becomes,

$$J(x+dx) - J(x) = \frac{\partial}{\partial x}J(x)dx + \frac{\partial^2}{\partial x^2}J(x)dx^2 \quad (49)$$

$$= \pm\frac{\partial}{\partial x}J(x)cdt + \frac{\partial^2}{\partial x^2}J(x)c^2dt^2 \quad (50)$$

$$(51)$$

for $dx^2 = c^2dt^2$. Setting $\frac{\partial}{\partial x}J(x) = 0$ means that the kernel is stable to the variation of data. This assumption gives the wave equation of the following form:

$$\frac{\partial^2}{\partial t^2}J(x) - c^2\frac{\partial^2}{\partial x^2}J(x) = 0. \quad (52)$$

The solution of this equation is

$$J(x) = \exp(i(ct \pm x)). \quad (53)$$

The objective function becomes the following form:

$$L_D(\alpha(t)) = \sum_{i=1}^{l}\alpha_i(t) - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i(t)\alpha_j(t+\Delta t)y_iy_\mathrm{J}J(x_i(t), x_j(t+\Delta t)), \quad (54)$$

here,

$$J(x_i(t), x_j(t+\Delta t)) = \exp[i(c\Delta t \pm \{x_j(t+\Delta t) - x_i(t)\})]. \quad (55)$$

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1624v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2016, publ: 5 Jan 2016

5 Comments

5.1 State Dependence

The assumption of flat space in higher dimension is playing a central role in the algorithms of SVM. The effectiveness of SVM verifies this assumption. In general, the spaces are not flat. If it remain in higher dimension, we have to introduce curved space. This is realized as state dependent coefficients in our framework. For stochastic case, this is represented as

$$dx = \mu(x,t)dt + \sigma(x,t)dz, \tag{56}$$

and for deterministic case,

$$dx^2 = c(x,t)^2 dt^2. \tag{57}$$

For these cases, we have to solve non-linear diffusion or wave equations. This needs obtaining the solutions, numerically. But this will realize a new mechanism by merging solving the equations in numerical way to algorithms.

5.2 Reinterpretation of Kernel

5.2.1 Kernel in Information Geometry

We consider the infinitismal distance between the points in data space.

$$ds^2 = |s(x+dx) - s(x)|^2 = \sum \left\{ \frac{\partial}{\partial x_i} s(x) \cdot \frac{\partial}{\partial x_j} s(x) \right\} dx_i dx_j \tag{58}$$

The metric of this space is

$$g_{ij}(x) = \left(\frac{\partial}{\partial x_i} s(x) \right) \cdot \left(\frac{\partial}{\partial x_j} s(x) \right). \tag{59}$$

This metric is represented with kernel as [7, 8]

$$g_{ij}(x) = \frac{\partial^2}{\partial x_i \partial x'_j} K(x,x')|_{x'=x}. \tag{60}$$

If we take the Gaussian kernel, $K(x,x')$ is

$$K(x,x') = \exp\left\{ -\frac{|x-x'|^2}{\sigma^2} \right\}. \tag{61}$$

This formalism can be extended to the time-dependent case. This is realized as the inclusion of time direction to the metric.

$$g_{ij} \mapsto g_{\mu\nu}, \tag{62}$$

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1624v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2016, publ: 5 Jan 2016

here,

$$\mu, \nu = 0, 1, \ldots, N. \tag{63}$$

This picture corresponds to deterministic case in our work.

5.2.2 Equivalence Principle in Information Geometry

Once include the time direction, we can deal the information space by the same way as in general relativity. Here, we consider a relationship that is a counterpart of equivalence principle [9]. We assume the small deviation from Minkowskian metric for the metric $g_{\mu\nu}$, in short, $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$, $h_{\mu\nu} \ll 1$. We can derive the following relation for the points of data,

$$\ddot{x}_\mu + \Gamma^\mu_{\alpha\beta} \dot{x}^\alpha \dot{x}^\beta = 0 \tag{64}$$

In the above equation, $(\dot{})$ denotes the derivative with respect to time, and $\Gamma^\mu_{\alpha\beta}$ is Christoffel symbol. For the space components, the following equation holds,

$$\ddot{x}_\mu = -\Gamma^\mu_{\alpha\beta} \dot{x}^\alpha \dot{x}^\beta \simeq -\Gamma^i_{00} c^2. \tag{65}$$

The last equality holds under the assumption of the very slow speed of point compared with the speed of light. The Christoffel symbol can be calculated as,

$$\Gamma^i_{00} \simeq -\frac{1}{2} \partial^i h_{00}. \tag{66}$$

Then we obtain the equation of following form,

$$\ddot{x}^i \simeq \frac{c^2}{2} \partial^i h_{00}. \tag{67}$$

The comparison of this equation with the equation of Newtonian mechanics: $\ddot{x}^i = -\partial^i \phi$, gives the following relation between potential and metric,

$$\phi \simeq -\frac{c^2}{2} h_{00}. \tag{68}$$

This gives the relation between the acceleration of point of data and the metric: dynamics and shape of data.

5.2.3 Relation to AdS/CFT Correspondence

In the context of AdS/CFT correspondence, diffusion equation and wave equation appear as a reflection of Fick's law [10, 11, 12]. The solutions of these

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1624v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2016, publ: 5 Jan 2016

equations are interpreted as kernel, described in former sections. In information geometry, kernel is related with metric of space as shown in above. It will be interesting to seek the interpretation of kernel in the context of AdS/CFT correspondence and to find the interpretation of AdS/CFT in the context of information geometry.

6 Conclusions

In this paper, we gave a derivation of kernel of DSVM. The time-dependent processes that we adopted were stochastic and deterministic. The derivation of kernels followed Bellman principle. We also gave physical interpretations in the context of extension of information geometry.

The information geometric comments that we gave are little, but the study of information geometry with physical viewpoints will be interesting and bring new insights both in the contexts of information theory and physics. We will explore them.

Followings are in the scope our future study, application to financial optimization problem, extension to dissipative systems, relation to heat equations in Topological Data Analysis, stochastic quantization of gravity, information paradox, information causality, and conformal transformation.

Acknowledgements

We greatly thank to kind hospitality of my colleagues for giving ideas on business issues and comfortable environments.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1624v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2016, publ: 5 Jan 2016

References

[1] Vapnik, V., Support-vector networks, Machine Learning 20 273 (1995).

[2] Krasotkina, O. V., Mottl, V. V., and Turkov, P. A., Bayesian Approach to the Pattern Recognition Problem in Nonstationary Environment, PReMI 2011, LNCS 6744, pp. 24-29.

[3] Guangzhi, S. D., Lianglong, H. J., and Yanxia, Z., Dynamic support vector machine by distributing kernel function, In Advanced Computer Control (ICACC), 2010 2nd International Conference on, vol. 2, pp. 362-365.

[4] Hong, W. C., Intelligent Energy Demand Forecasting, Springer (2013).

[5] Bishop, C. M., Pattern Recognition and Machine Learning, Springer (2006).

[6] Chang, F. R., Stochastic Optimization in Continuous Time, Cambridge (2004).

[7] Amari, S., Nagaoka, H., and Harada, D., Methods of Information Geometry, AMS (2007).

[8] Amari, S., New developments in information geometry, Sience-sha (2014), in Japanese.

[9] Landau, L. D. and Lifshitz, E. M., The Classical Theory of Fields, Butterworth-Heinemann (1980).

[10] Natsuume, M., AdS/CFT Duality Use Guide, Springer (2015).

[11] Ammon, M. and Erdmenger, J., Gauge/Gravity Duality, Cambridge (2015).

[12] Nstase, H., Introduction to the AdS/CFT Correspondence, Cambridge (2015).

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1624v1 | CC-BY 4.0 Open Access | rec: 5 Jan 2016, publ: 5 Jan 2016