Appendix for "A Kernelized Stein Discrepancy for Goodness-of-fit Tests"

Qiang Liu QLIU@CS.DARTMOUTH.EDU

Computer Science, Dartmouth College, NH, 03755

Jason D. Lee JASONDLEE88@EECS.BERKELEY.EDU

Michael Jordan JORDAN@CS.BERKELEY.EDU

Department of Electrical Engineering and Computer Science University of California, Berkeley, CA 94709

A. Proofs

Proof of Theorem 3.6. 1) Denote by $\boldsymbol{v}(x, x') = k(x, x')\boldsymbol{s}_q(x') + \nabla_{x'} k(x, x') = \mathcal{A}_q k_x(x')$; applying Lemma 2.3 on $k(x, \cdot)$ with fixed x,

$$\begin{aligned}\mathbb{S}(p, q) &= \mathbb{E}_{x,x'\sim p}[(\boldsymbol{s}_q(x) - \boldsymbol{s}_p(x))^\top k(x, x')(\boldsymbol{s}_q(x') - \boldsymbol{s}_p(x'))] \\ &= \mathbb{E}_{x,x'\sim p}[(\boldsymbol{s}_q(x) - \boldsymbol{s}_p(x))^\top \boldsymbol{v}(x, x')]\end{aligned}$$

Because $k(\cdot, x')$ is in the Stein class of p for any x', we can show that $\nabla_{x'} k(\cdot, x')$ is also in the Stein class, since

$$\int_x \nabla_x (p(x)\nabla_{x'} k(x, x'))dx = \nabla_{x'} \int_x \nabla_x (p(x)k(x, x'))dx = 0,$$

and hence $\boldsymbol{v}(\cdot, x')$ is also in the Stein class; apply Lemma 2.3 on $\boldsymbol{v}(\cdot, x')$ with fixed x' gives

$$\begin{aligned}\mathbb{S}(p, q) &= \mathbb{E}_{x,x'\sim p}[(\boldsymbol{s}_q(x) - \boldsymbol{s}_p(x))^\top \boldsymbol{v}(x, x'))] \\ &= \mathbb{E}_{x,x'\sim p}[\boldsymbol{s}_q(x)^\top \boldsymbol{v}(x, x') + \mathrm{trace}(\nabla_x \boldsymbol{v}(x, x'))]\end{aligned}$$

The result then follows by noting that $\nabla_x \boldsymbol{v}(x, x') = \nabla_x k(x, x')\boldsymbol{s}_q(x')^\top + \nabla_{x'x'} k(x, x')$. □

Proof of Theorem 3.7. Note that

$$\nabla_x k(x, x') = \sum_j \lambda_j \nabla_x e_j(x)\, e_j(x'), \qquad \nabla_{x,x'} k(x, x') = \sum_j \lambda_j \nabla_x e_j(x)\, \nabla_{x'} e_j(x')^\top,$$

and hence

$$\begin{aligned}&u_q(x, x') \\ &= \boldsymbol{s}_q(x)^\top k(x, x')\boldsymbol{s}_q(x') + \boldsymbol{s}_q(x)^\top \nabla'_x k(x, x') + \boldsymbol{s}_q(x')^\top \nabla_x k(x, x') + \mathrm{trace}(\nabla_{x,x'} k(x, x') \\ &= \sum_j \lambda_j \left[\boldsymbol{s}_q(x)^\top e_j(x)e_j(x')\boldsymbol{s}_q(x') + \boldsymbol{s}_q(x)^\top e_j(x)\nabla_{x'} e_j(x') + \boldsymbol{s}_q(x')^\top \nabla_x e_j(x)e_j(x') + \nabla_x e_j(x)^\top \nabla_{x'} e_j(x')\right] \\ &= \sum_j \lambda_j \left[\boldsymbol{s}_q(x)e_j(x) + \nabla_x e_j(x)\right]^\top \left[\boldsymbol{s}_q(x')e_j(x') + \nabla_{x'} e_j(x')\right] \\ &= \sum_j \lambda_j [\mathcal{A}_q e_j(x)]^\top [\mathcal{A}_q e_j(x')].\end{aligned}$$

Therefore, $u_q(x, x')$ is positive definite because $\lambda_j > 0$. In addition,

$$\begin{aligned}\mathbb{S}(p,q) &= \mathbb{E}_{x,x'}[u_q(x,x')] \\ &= \sum_j \lambda_j \mathbb{E}_x[\mathcal{A}_q e_j(x)]^\top \, \mathbb{E}_{x'}[\mathcal{A}_q e_j(x')] \\ &= \sum_j \lambda_j ||\mathbb{E}_x[\mathcal{A}_q e_j(x)]||_2^2.\end{aligned}$$

□

Proof of Theorem 3.8. We first prove (12) by applying the reproducing property $k(x,x') = \langle k(x,\cdot), k(x',\cdot)\rangle_{\mathcal{H}}$ on (8):

$$\begin{aligned}\mathbb{S}(p,q) &= \mathbb{E}_{x,x'\sim p}[(\boldsymbol{s}_q(x) - \boldsymbol{s}_p(x))^\top \; k(x,x') \; (\boldsymbol{s}_q(x') - \boldsymbol{s}_p(x'))] \\ &= \mathbb{E}_{x,x'\sim p}[(\boldsymbol{s}_q(x) - \boldsymbol{s}_p(x))^\top \; \big\langle k(x,\cdot), \; k(x,\cdot)\big\rangle_{\mathcal{H}} \; (\boldsymbol{s}_q(x') - \boldsymbol{s}_p(x'))] \\ &= \sum_{\ell=1}^d \big\langle \mathbb{E}_x[(\boldsymbol{s}_q^\ell(x) - \boldsymbol{s}_p^\ell(x))k(x,\cdot)], \; \mathbb{E}_{x'}[k(x,\cdot)(\boldsymbol{s}_q^\ell(x) - \boldsymbol{s}_p^\ell(x))]\big\rangle_{\mathcal{H}} \\ &= \sum_{\ell=1}^d \big\langle \boldsymbol{\beta}_\ell, \boldsymbol{\beta}_\ell \big\rangle_{\mathcal{H}} \\ &= ||\boldsymbol{\beta}||_{\mathcal{H}^d}^2\end{aligned}$$

where we used the fact that $\boldsymbol{\beta}(x') = \mathbb{E}_{x\sim p}[\mathcal{A}_q k_{x'}(x)] = \mathbb{E}_{x\sim p}[(\boldsymbol{s}_q(x)k(x,x') + \nabla_x k(x,x')] = \mathbb{E}_x[(\boldsymbol{s}_q(x) - \boldsymbol{s}_p(x))k(x,x')]$. In addition,

$$\begin{aligned}\langle \boldsymbol{f}, \boldsymbol{\beta}\rangle_{\mathcal{H}^d} &= \sum_{\ell=1}^d \langle f_\ell, \; \mathbb{E}_{x\sim p}[(\boldsymbol{s}_q^\ell(x)k(x,\cdot) + \nabla_{x_\ell} k(x,\cdot)]\rangle_{\mathcal{H}} \\ &= \sum_{\ell=1}^d \mathbb{E}_{x\sim p}[(\boldsymbol{s}_q^\ell(x)\langle f_\ell, k(x,\cdot)\rangle_{\mathcal{H}} + \langle f_\ell, \nabla_{x_\ell} k(x,\cdot)\rangle_{\mathcal{H}}] \\ &= \sum_{\ell=1}^d \mathbb{E}_{x\sim p}[(\boldsymbol{s}_q^\ell(x) f_\ell(x) + \nabla_{x_\ell} f_\ell(x)] \\ &= \mathbb{E}_{x\sim p}[\mathrm{trace}(\mathcal{A}_q \boldsymbol{f}(x))],\end{aligned}$$

where we used the fact that $\nabla_x f(x) = \langle f(\cdot), \; \nabla_x k(x,\cdot)\rangle_{\mathcal{H}}$; see (Zhou, 2008; Steinwart & Christmann, 2008). The variational form (13) then follows the fact that $||\boldsymbol{\beta}||_{\mathcal{H}^d} = \max_{\boldsymbol{f}\in\mathcal{H}^d} \big\{\langle \boldsymbol{f}, \boldsymbol{\beta}\rangle_{\mathcal{H}^d}, \quad .s.t. \;\; ||\boldsymbol{f}||_{\mathcal{H}^d} \leq 1\big\}$.

Finally, the $\boldsymbol{\beta}(\cdot) = \mathbb{E}_{x\sim p}[(\boldsymbol{s}_q(x)k(x,\cdot) + \nabla_x k(x,\cdot)]$ is in the Stein class of p because $k(x,\cdot)$ and $\nabla_x k(x,\cdot)$ are in the Stein class of p for any fixed x (see the proof of Theorem 3.6). □

Proof Proposition 3.5. For any $f \in \mathcal{H}$ with kernel $k(x,x')$, we have $f = \langle f, \; k(\cdot,x)\rangle_{\mathcal{H}}$ and $\nabla_x f = \langle f, \; \nabla_x k(x,\cdot)\rangle_{\mathcal{H}}$. Therefore,

$$\begin{aligned}\mathbb{E}_{x\sim p}[\boldsymbol{s}_p(x)f(x) + \nabla_x f(x)] &= \mathbb{E}_{x\sim p}[\boldsymbol{s}_p(x)\big\langle f, \; k(x,\cdot)\big\rangle_{\mathcal{H}} + \big\langle f, \; \nabla_x k(x,\cdot)\big\rangle_{\mathcal{H}}] \\ &= \big\langle f, \;\; \mathbb{E}_{x\sim p}[\boldsymbol{s}_p(x)k(x,\cdot) + \nabla_x k(x,\cdot)]\big\rangle_{\mathcal{H}} \\ &= \big\langle f, \;\; \mathbb{E}_{x\sim p}[\mathcal{A}_p k_x(\cdot)]\rangle_{\mathcal{H}} \\ &= 0,\end{aligned}$$

where the last step used the fact that $\mathbb{E}_{x\sim p}[\mathcal{A}_p k_x(\cdot)]$ because $k_x(\cdot) = k(\cdot, x)$ is in the Stein class of p for any fixed x. □

Proof of Theorem 4.1. Applying the standard asymptotic results of U-statistics in Serfling (2009, Section 5.5), we just need to check that $\sigma_u^2 \neq 0$ when $p \neq q$ and $\sigma_u^2 = 0$ when $p = q$.

We first note that we can show that $\mathbb{E}_{x'\sim p}[u_q(x,x')] = \mathrm{trace}(\mathcal{A}_q \boldsymbol{\beta})$, where $\boldsymbol{\beta}(x) = \mathbb{E}_{x'\sim p}[\mathcal{A}_q k_x(x')]$ and is in the Stein class of p (see the proof of Theorem 3.6). Therefore, when $p = q$, we have $\boldsymbol{\beta}(x) \equiv 0$ by Stein's identity, and hence $\sigma_u^2 = 0$. Assume $\sigma_u^2 = 0$ when $p \neq q$, we must have $\mathbb{E}_{x'\sim p}[u_q(x,x')] = c$, where c is a constant. Therefore,

$$c = \mathbb{E}_{x\sim q}\big(\mathbb{E}_{x'\sim p}[u_q(x,x')]\big) = \mathbb{E}_{x'\sim p}\big(\mathbb{E}_{x\sim q}[u_q(x,x')]\big).$$

Because we can show that $\mathbb{E}_{x\sim q}[u_q(x,x')] = 0$ following the proof above for $p = q$, we must have $c = 0$, and hence

$$\mathbb{S}(p,q) = \mathbb{E}_{x\sim p}\big(\mathbb{E}_{x'\sim p}[u_q(x,x')]\big) = c = 0,$$

which contradicts with $p \neq q$.

□

Proof of Theorem 5.1. (19) is obtained by applying Cauchy-Schwarz inequality on (8),

$$\begin{aligned}
\mathbb{S}(p,q)^2 &= |\mathbb{E}_{xx'}[(\boldsymbol{s}_q(x) - \boldsymbol{s}_p(x))^\top k(x,x')(\boldsymbol{s}_q(x) - \boldsymbol{s}_p(x))]|^2 \\
&\leq \mathbb{E}_{xx'}[k(x,x')^2] \cdot\ \mathbb{E}_{x,x'}[[(\boldsymbol{s}_q(x) - \boldsymbol{s}_p(x))^\top(\boldsymbol{s}_q(x') - \boldsymbol{s}_p(x'))]^2] \\
&\leq \mathbb{E}_{xx'}[k(x,x')^2] \cdot\ \mathbb{E}_{x,x'}[||\boldsymbol{s}_q(x) - \boldsymbol{s}_p(x)||_2^2 \cdot ||\boldsymbol{s}_q(x) - \boldsymbol{s}_p(x)||_2^2] \\
&= \mathbb{E}_{xx'}[k(x,x')^2] \cdot \mathbb{F}(p,q)^2.
\end{aligned}$$

To prove (20), we simply note that (13) is equivalent to

$$\sqrt{\mathbb{S}(p,q)} = \max_{\boldsymbol{f}\in\mathcal{H}^d}\Big\{\mathbb{E}_p[(\boldsymbol{s}_q(x) - \boldsymbol{s}_p(x))^\top \boldsymbol{f}(x)] \quad s.t. \quad ||\boldsymbol{f}||_{\mathcal{H}^d} \leq 1\Big\}.$$

Taking $\boldsymbol{f} = (\boldsymbol{s}_q - \boldsymbol{s}_p)/||\boldsymbol{s}_q(x) - \boldsymbol{s}_p(x)||_{\mathcal{H}^d}$ then gives (20). □

Proposition A.1. *Let $\mathcal{F}(p) = \mathcal{L}^2(p) \cap \mathcal{S}(p)$, where $\mathcal{S}(p)$ represents the Stein class of p, then we have*

$$\sqrt{\mathbb{F}(p,q)} \geq \max_{\boldsymbol{f}\in\mathcal{F}(p)^d}\Big\{\mathbb{E}_p[\mathrm{trace}(\mathcal{A}_q\boldsymbol{f}(x))] \quad s.t. \quad \mathbb{E}_p[||\boldsymbol{f}(x)||_2^2] \leq 1\Big\}.$$

and the equality holds when $\boldsymbol{s}_q - \boldsymbol{s}_p \in \mathcal{F}(p)^d$.

Note that $\mathcal{L}^2(p)$ is larger than the Stein class and RKHS, and includes discontinuous, non-smooth functions, and hence we need to ensure $\boldsymbol{f}$ is in the Stein class explicitly.

Proof. Denote by $(\mathcal{L}^2(p))^d = \mathcal{L}^2(p) \times \cdots \times \mathcal{L}^2(p)$, note that by the definition of $\mathbb{F}(p,q)$, we have

$$\sqrt{\mathbb{F}(p,q)} = \max_{\boldsymbol{f}\in(\mathcal{L}^2(p))^d}\Big\{\sum_{\ell=1}^d \mathbb{E}_p[f_\ell(x)(\boldsymbol{s}_q^\ell(x) - \boldsymbol{s}_p^\ell(x))] \quad s.t. \quad \mathbb{E}_p[||\boldsymbol{f}(x)||_2^2] \leq 1\Big\}. \tag{A.1}$$

Restricting the maximizing to $\mathcal{F}(p)^d$ and applying Lemma 2.3 would give the result. □

References

Arcones, M. A. and Gine, E. On the bootstrap of U and V statistics. *The Annals of Statistics*, pp. 655–674, 1992.

Birge, L. and Massart, P. Estimation of integral functionals of a density. *The Annals of Statistics*, pp. 11–29, 1995.

Chandrasekaran, V., Srebro, N., and Harsha, P. Complexity of inference in graphical models. In *UAI*. July 2008.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. *Monte Carlo methods in Bayesian computation*. Springer Science & Business Media, 2012.

Chwialkowski, K. P., Sejdinovic, D., and Gretton, A. A wild bootstrap for degenerate kernel tests. In *Advances in neural information processing systems*, pp. 3608–3616, 2014.

Gorham, J. and Mackey, L. Measuring sample quality with stein's method. In *NIPS*, pp. 226–234, 2015.

Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. A fast, consistent kernel two-sample test. In *Advances in neural information processing systems*, pp. 673–681, 2009.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Hall, P. and Marron, J. S. Estimation of integrated squared density derivatives. *Statistics & Probability Letters*, 6(2): 109–115, 1987.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science*, 313(5786): 504–507, 2006.

Ho, H.-C. and Shieh, G. S. Two-stage U-statistics for hypothesis testing. *Scandinavian journal of statistics*, 33(4):861–873, 2006.

Hoeffding, W. A class of statistics with asymptotically normal distribution. *The annals of mathematical statistics*, pp. 293–325, 1948.

Huskova, M. and Janssen, P. Consistency of the generalized bootstrap for degenerate U-statistics. *The Annals of Statistics*, pp. 1811–1823, 1993.

Hyvärinen, A. Estimation of non-normalized statistical models by score matching. In *Journal of Machine Learning Research*, pp. 695–709, 2005.

Johnson, O. *Information theory and the central limit theorem*, volume 8. World Scientific, 2004.

Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Krishnamurthy, A., Kandasamy, K., Poczos, B., and Wasserman, L. Nonparametric estimation of Renyi divergence and friends. In *ICML*, 2014.

Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

Ley, C. and Swan, Y. Stein's density approach and information inequalities. *Electron. Comm. Probab*, 18(7):1–14, 2013.

Lyu, S. Interpretation and generalization of score matching. In *UAI*, pp. 359–366, 2009.

Marsden, J. E. and Tromba, A. *Vector calculus*. Macmillan, 2003.

Neal, R. M. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Salakhutdinov, R. Learning deep generative models. *Annual Review of Statistics and Its Application*, 2(1):361–385, 2015.

Salakhutdinov, R. and Murray, I. On the quantitative analysis of deep belief networks. In *ICML*, pp. 872–879, 2008.

Serfling, R. J. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

Sriperumbudur, B., Fukumizu, K., Kumar, R., Gretton, A., and Hyvärinen, A. Density estimation in infinite dimensional exponential families. *arXiv preprint arXiv:1312.3516*, 2013.

Stein, C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, pp. 583–602, 1972.

Stein, C., Diaconis, P., Holmes, S., Reinert, G., et al. Use of exchangeable pairs in the analysis of simulations. In *Stein's Method*, pp. 1–25. Institute of Mathematical Statistics, 2004.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Welling, M., Rosen-Zvi, M., and Hinton, G. E. Exponential family harmoniums with an application to information retrieval. In *Advances in neural information processing systems*, pp. 1481–1488, 2004.

Zaremba, W., Gretton, A., and Blaschko, M. B. B-tests: Low variance kernel two-sample tests. In *NIPS*, pp. 755–763, 2013.

Zhou, D.-X. Derivative reproducing properties for kernel methods in learning theory. *Journal of computational and Applied Mathematics*, 220(1):456–463, 2008.