Supplementary Material to Learning from Distributions via Support Measure Machines

Krikamol Muandet
MPI for Intelligent Systems, Tübingen
krikamol@tuebingen.mpg.de

Kenji Fukumizu
The Institute of Statistical Mathematics, Tokyo
fukumizu@ism.ac.jp

Francesco Dinuzzo
MPI for Intelligent Systems, Tübingen
fdinuzzo@tuebingen.mpg.de

Bernhard Schölkopf
MPI for Intelligent Systems, Tübingen
bs@tuebingen.mpg.de

1 Proof of Theorem 1

Theorem 1. *Given training examples $(\mathbb{P}_i, y_i) \in \mathscr{P} \times \mathbb{R}$, $i = 1, \dots, m$, a strictly monotonically increasing function $\Omega : [0, +\infty) \to \mathbb{R}$, and a loss function $\ell : (\mathscr{P} \times \mathbb{R}^2)^m \to \mathbb{R} \cup \{+\infty\}$, any $f \in \mathcal{H}$ minimizing the regularized risk functional*

$$\ell\left(\mathbb{P}_1, y_1, \mathbb{E}_{\mathbb{P}_1}[f], \dots, \mathbb{P}_m, y_m, \mathbb{E}_{\mathbb{P}_m}[f]\right) + \Omega\left(\|f\|_{\mathcal{H}}\right) \tag{1}$$

admits a representation of the form $f = \sum_{i=1}^m \alpha_i \mu_{\mathbb{P}_i}$ for some $\alpha_i \in \mathbb{R}$, $i = 1, \dots, m$.

Proof. By virtue of Proposition 2 in [1], the linear functional $\mathbb{E}_{\mathbb{P}}[\cdot]$ are bounded for all $\mathbb{P} \in \mathscr{P}$. Then, given $\mathbb{P}_1, \mathbb{P}_2, ..., \mathbb{P}_m$, any $f \in \mathcal{H}$ can be decomposed as

$$f = f_\mu + f^\perp$$

where $f_\mu \in \mathcal{H}$ lives in the span of $\mu_{\mathbb{P}_i}$, i.e., $f_\mu = \sum_{i=1}^m \alpha_i \mu_{\mathbb{P}_i}$ and $f^\perp \in \mathcal{H}$ satisfying, for all j, $\langle f^\perp, \mu_{\mathbb{P}_j} \rangle = 0$. Hence, for all j, we have

$$\mathbb{E}_{\mathbb{P}_j}[f] = \mathbb{E}_{\mathbb{P}_j}[f_\mu + f^\perp] = \langle f_\mu + f^\perp, \mu_{\mathbb{P}_j} \rangle = \langle f_\mu, \mu_{\mathbb{P}_j} \rangle + \langle f^\perp, \mu_{\mathbb{P}_j} \rangle = \langle f_\mu, \mu_{\mathbb{P}_j} \rangle$$

which is independent of $f^\perp$. As a result, the loss functional ℓ in (1) does not depend on $f^\perp$. For the regularization functional Ω, since $f^\perp$ is orthogonal to $\sum_{i=1}^m \alpha_i \mu_{\mathbb{P}_i}$ and Ω is strictly monotonically increasing, we have

$$\Omega(\|f\|) = \Omega(\|f_\mu + f^\perp\|) = \Omega(\sqrt{\|f_\mu\|^2 + \|f^\perp\|^2}) \geq \Omega(\|f_\mu\|)$$

with equality if and only if $f^\perp = 0$ and thus $f = f_\mu$. Consequently, any minimizer must take the form $f = \sum_{i=1}^m \alpha_i \mu_{\mathbb{P}_i} = \sum_{i=1}^m \alpha_i \mathbb{E}_{\mathbb{P}_i}[k(x, \cdot)]$. ∎

2 Proof of Theorem 3

Theorem 3. *Given an arbitrary probability distribution $\mathbb{P}$ with variance σ^2, a Lipschitz continuous function $f : \mathbb{R} \to \mathbb{R}$ with constant C_f, an arbitrary loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ that is Lipschitz continuous in the second argument with constant C_ℓ, it follows that*

$$|\mathbb{E}_{x\sim\mathbb{P}}[\ell(y, f(x))] - \ell(y, \mathbb{E}_{x\sim\mathbb{P}}[f(x)])| \leq 2C_\ell C_f \sigma$$

for any $y \in \mathbb{R}$.

Proof. Assume that x is distributed according to $\mathbb{P}$. Let m_X be the mean of X in $\mathbb{R}^d$. Thus, we have

$$\begin{aligned}
|\mathbb{E}_{\mathbb{P}}[\ell(y,f(x))]-\ell(y,\mathbb{E}_{\mathbb{P}}[f(x)])| &\leq \int |\ell(y,f(\tilde{x}))-\ell(y,\mathbb{E}_{\mathbb{P}}[f(x)])|\,\mathrm{d}\mathbb{P}(\tilde{x}) \\
&\leq C_\ell \int |f(\tilde{x})-\mathbb{E}_{\mathbb{P}}[f(x)]|\,\mathrm{d}\mathbb{P}(\tilde{x}) \\
&\leq \underbrace{C_\ell \int |f(\tilde{x})-f(m_X)|\,\mathrm{d}\mathbb{P}(\tilde{x})}_{A} + \underbrace{C_\ell|f(m_X)-\mathbb{E}_{\mathbb{P}}[f(x)]|}_{B} \ .
\end{aligned}$$

Control of (A) The first term is upper bounded by

$$C_\ell \int C_f\|\tilde{x}-m_X\|\,\mathrm{d}\mathbb{P}(\tilde{x}) \leq C_\ell C_f \sigma \ \ , \tag{2}$$

where the last inequality is given by $\mathbb{E}_{\mathbb{P}}[\|\tilde{x}-m_X\|] \leq \sqrt{\mathbb{E}_{\mathbb{P}}[\|\tilde{x}-m_X\|^2]} = \sigma$.

Control of (B) Similarly, the second term is upper bounded by

$$C_\ell \left| \int f(m_X)-f(\tilde{x}) \right| \mathrm{d}\mathbb{P}(\tilde{x}) \leq C_\ell \int C_f \|m_X-\tilde{x}\|\,\mathrm{d}\mathbb{P}(\tilde{x}) \leq C_\ell C_f \sigma \ \ . \tag{3}$$

Combining (2) and (3) yields

$$|\mathbb{E}_{\mathbb{P}}[\ell(y,f(x))]-\ell(y,\mathbb{E}_{\mathbb{P}}[f(x)])| \leq 2C_\ell C_f \sigma \ \ ,$$

thus completing the proof. ■

3 Proof of Lemma 4

Lemma 4. *Let $k(x,z)$ be a bounded p.d. kernel on a measure space such that $\iint k(x,z)^2\,\mathrm{d}x\,\mathrm{d}z < \infty$, and $g(x,\tilde{x})$ be a square integrable function such that $\int g(x,\tilde{x})\,\mathrm{d}\tilde{x} < \infty$ for all x. Given a sample $\{(\mathbb{P}_i,y_i)\}_{i=1}^m$ where each $\mathbb{P}_i$ is assumed to have a density given by $g(x_i,x)$, the linear SMM is equivalent to the SVM on the training sample $\{(x_i,y_i)\}_{i=1}^m$ with kernel $K_g(x,z) = \iint k(\tilde{x},\tilde{z})g(x,\tilde{x})g(z,\tilde{z})\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{z}$.*

Proof. For a training sample $\{(x_i,y_i)\}_{i=1}^m$, the SVM with kernel K_g minimizes

$$\ell(\{x_i,y_i,f(x_i)+b\}_{i=1}^m) + \lambda\|f\|^2_{\mathcal{H}_{K_g}} \ \ .$$

By the representer theorem, $f(x)=\sum_{i=1}^m \alpha_i K_g(x,x_j)$ with some $\alpha_i \in \mathbb{R}$, hence this is equivalent to

$$\ell(\{x_i,y_i,\sum_{j=1}^m \alpha_j K_g(x_i,x_j)+b\}_{i=1}^m) + \lambda \sum_{i,j=1}^m \alpha_i\alpha_j K_g(x_i,x_j) \ \ .$$

Next, consider the kernel mean of the probability measure $g(x_i,x)dx$ given by $\mu_i = \int k(\cdot,\tilde{x})g(x_i,\tilde{x})\,\mathrm{d}\tilde{x}$ and note that $\langle \mu_i, f\rangle_{\mathcal{H}_k} = \int f(\tilde{x})g(x_i,\tilde{x})\,\mathrm{d}\tilde{x}$ for any $f \in \mathcal{H}_k$. The linear SMM with loss ℓ and kernel k minimizes

$$\ell(\{\mathbb{P}_i,y_i,\langle\mu_i,f\rangle_{\mathcal{H}_k}+b\}_{i=1}^m) + \lambda\|f\|^2_{\mathcal{H}_k} \ \ .$$

By Theorem 1, each minimizer f admits a representation of the form

$$f = \sum_{j=1}^m \alpha_j\mu_j = \sum_{j=1}^m \alpha_j \int k(\cdot,\tilde{x})g(x_j,\tilde{x})\,\mathrm{d}\tilde{x} \ \ .$$

Thus, for this f we have

$$\langle\mu_i,f\rangle_{\mathcal{H}_k} = \sum_{j=1}^m \alpha_j \iint k(\tilde{z},\tilde{x})g(x_i,\tilde{x})g(x_j,\tilde{z})\,\mathrm{d}\tilde{x}\,\mathrm{d}\tilde{z} = \sum_{j=1}^m \alpha_j K_g(x_i,x_j)$$

and

$$\|f\|^2_{\mathcal{H}_k} = \sum_{i,j=1}^m \alpha_i\alpha_j\langle\mu_i,\mu_j\rangle = \sum_{i,j=1}^m \alpha_i\alpha_j K_g(x_i,x_j)$$

, as above. This completes the proof. ■

References

[1] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 99:1517–1561, 2010.