Supplement to "Finite sample properties of parametric MMD estimation: robustness to misspecification and dependence"

BADR-EDDINE CHÉRIEF-ABDELLATIF[1,*] and PIERRE ALQUIER[2,**]

[1] *Department of Statistics, University of Oxford.*
E-mail: [*]`badr-eddine.cherief-abdellatif@stats.ox.ac.uk`

[2] *RIKEN AIP, Tokyo, Japan.*
E-mail: [**]`pierrealain.alquier@riken.jp`

This is the supplement to [Chérief-Abdellatif and Alquier, 2021]. It contains the proofs of the results stated in Sections 4 and 5 of this paper.

As five equations are numbered in the paper, we start here the numbering of equations at (6), so that it is clear that for example (3) refers to Equation 3 in the paper.

Appendix A: Proofs of Section 4

Proof of Proposition 4.1. We remind that $P_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$ where $\theta \in \Theta = \mathbb{R}^d$. When X and Y are independent, respectively from P_θ and $P_{\theta'}$, we have $(X - Y) \sim \mathcal{N}(\theta - \theta', \sigma^2 I_d)$. Thus,

$$\frac{(X-Y)}{\sqrt{2\sigma^2}} \sim \mathcal{N}\left(\frac{(\theta-\theta')}{\sqrt{2\sigma^2}}, I_d\right)$$

and thus the square of this random variable is a noncentral chi-square random variable:

$$\frac{\|X-Y\|^2}{2\sigma^2} \sim \chi^2\left(d, \frac{\|\theta-\theta'\|^2}{2\sigma^2}\right).$$

It is known that when $U \sim \chi^2(d,m)$ we have $\mathbb{E}[\exp(tU)] = \exp(mt/(1-2t))/(1-2t)^{d/2}$. Taking $t = -(2\sigma^2)/\gamma^2$, this leads to

$$\left\langle \mu_{P_\theta}, \mu_{P_{\theta'}} \right\rangle_{\mathcal{H}} = \mathbb{E}_{X\sim P_\theta, Y\sim P_{\theta'}}\left[\exp\left(-\frac{\|X-Y\|^2}{\gamma^2}\right)\right] = \left(\frac{\gamma^2}{4\sigma^2+\gamma^2}\right)^{\frac{d}{2}} \exp\left(-\frac{\|\theta-\theta'\|^2}{4\sigma^2+\gamma^2}\right)$$

and thus

$$\mathbb{D}^2_{k_\gamma}(P_\theta, P_{\theta'}) = 2\left(\frac{\gamma^2}{4\sigma^2+\gamma^2}\right)^{\frac{d}{2}}\left[1-\exp\left(-\frac{\|\theta-\theta'\|^2}{4\sigma^2+\gamma^2}\right)\right]. \tag{6}$$

From (6) and Proposition 3.5 of the paper, we obtain, with probability at least $1-\delta$,

$$2\left(\frac{\gamma^2}{4\sigma^2+\gamma^2}\right)^{\frac{d}{2}}\left[1-\exp\left(-\frac{\|\theta-\tilde{\theta}_n\|^2}{4\sigma^2+\gamma^2}\right)\right] = \mathbb{D}^2_k\left(P_{\tilde{\theta}_n}, P_{\theta_0}\right) \leq 16\left(\varepsilon + \frac{1+\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}\right)^2,$$

that is,

$$\|\theta-\tilde{\theta}_n\|^2 \leq -(4\sigma^2+\gamma^2)\log\left\{1-8\left(\frac{4\sigma^2+\gamma^2}{\gamma^2}\right)^{\frac{d}{2}}\left(\varepsilon+\frac{1+\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}\right)^2\right\}.$$

Now, use inside the log the inequality $(1+1/x)^x \leq \mathrm{e}$ on $x = \gamma^2/(4\sigma^2)$ to get:

$$\|\theta - \tilde{\theta}_n\|^2 \leq -(4\sigma^2+\gamma^2)\log\left\{1 - 8\mathrm{e}^{\frac{2\sigma^2 d}{\gamma^2}}\left(\varepsilon + \frac{1+\sqrt{2\log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}\right)^2\right\}.$$

This proves the first inequality, that is, (2). Simply plug $\gamma = \sigma\sqrt{2d}$ to get the second inequality. □

The Gaussian example appears to be a very special case, where it is possible to derive an explicit formula for the MMD distance. In most models, this is not possible. However, in many models, we observed that it is actually possible to provide an explicit formula for the L^2 distance, of the form $\|\theta-\theta'\|^2 = F(\|p_\theta - p_{\theta'}\|^2_{L^2})$, where P_θ has a density $p_\theta \in L^2$ with respect to the Lebesgue measure and F is a nondecreasing function. It is then tempting to use the connection between the MMD distance and the L^2 distance mentioned in Remark 3.2 of the paper and in [Sriperumbudur et al., 2010]. The scheme of the proof is as follows:

$$\|\hat{\theta} - \theta_0\|^2 = F\left(\|p_{\hat{\theta}} - p_{\theta_0}\|^2_{L^2}\right) \underbrace{\leq}_{?} F\left(c\mathbb{D}^2_k(P_{\hat{\theta}}, P_{\theta_0})\right) \underbrace{\leq}_{\substack{\text{Theorem 3.1}\\ \text{Theorem 3.2}\\ \text{Proposition 3.5}\\ \dots}} F(\text{bound})$$

which means that in such models, the only new step to get a rate of convergence on $\hat{\theta}$ is to check the condition

$$\|p_\theta - p_{\theta'}\|^2_{L^2} \leq c\mathbb{D}^2_k(P_\theta, P_{\theta'}) \tag{7}$$

for some $c > 0$. Finally, in any case where $k_\gamma(x,x') = K(\|x-x'\|/\gamma)$, let μ denote the Fourier transform of K:

$$\mu(t) := \mathcal{F}[K](t) = \int K(x)\exp(-2i\pi\langle t, x\rangle)\mathrm{d}x.$$

Note that for the Gaussian kernel in $\mathbb{R}^d$, $K(u) = \exp(-\|u\|^2)$, $\mu(t) = \pi^{d/2}\exp(-\|t\|^2/4)$. We remind a few properties of the Fourier transform. First,

$$\mathcal{F}[K(\cdot/\gamma)](t) = \gamma^d \mathcal{F}[K](\gamma t) = \gamma^d \mu(\gamma t). \tag{8}$$

Let $\star$ denote the convolution product:

$$p \star q(x) = \int p(x-t)q(t)\mathrm{d}t$$

and we remind the classical result

$$\mathcal{F}[p \star q] = \mathcal{F}[p]\mathcal{F}[q]. \tag{9}$$

Finally, we remind that

$$\int p(x)\overline{q(x)}\mathrm{d}x = \int \mathcal{F}[p](t)\overline{\mathcal{F}[q](t)}\mathrm{d}t. \tag{10}$$

Then, we have:

$$\begin{aligned}
\mathbb{D}^2_{k_\gamma}(P_\theta, P_{\theta'}) &= \iint K(\|x-y\|/\gamma)[p_\theta(x) - p_{\theta'}(x)][p_\theta(y) - p_{\theta'}(y)]\mathrm{d}x\mathrm{d}y \\
&= \int [K(\cdot/\gamma) \star (p_\theta - p_{\theta'})](y)[p_\theta(y) - p_{\theta'}(y)]\mathrm{d}y, \text{ by definition of } \star \\
&= \int [K(\cdot/\gamma) \star (p_\theta - p_{\theta'})](y)\overline{[p_\theta(y) - p_{\theta'}(y)]}\mathrm{d}y, \text{ (densities are real-valued)} \\
&= \int \mathcal{F}[K(\cdot/\gamma) \star (p_\theta - p_{\theta'})](t)\overline{\mathcal{F}[p_\theta - p_{\theta'}](t)}\mathrm{d}t, \text{ by (10)}
\end{aligned}$$

$$
\begin{aligned}
&= \int \mathcal{F}[K(\cdot/\gamma)] \mathcal{F}[p_\theta - p_{\theta'}](t) \overline{\mathcal{F}[p_\theta - p_{\theta'}](t)} \mathrm{d}t, \text{ by (9)} \\
&= \int \gamma^d \mu(\gamma t) \left| \mathcal{F}[p_\theta - p_{\theta'}](t) \right|^2 \mathrm{d}t, \text{ by (8)}.
\end{aligned}
$$

So, an alternative way to check (7) is to check that:

$$
\| p_\theta - p_{\theta'} \|_{L^2}^2 \leq c \int \gamma^d \mu(\gamma t) \left| \mathcal{F}[p_\theta - p_{\theta'}](t) \right|^2 \mathrm{d}t.
$$

Proof of Proposition 4.2. We have

$$
\langle p_\theta, p_{\theta'} \rangle_{L^2} = \frac{2}{\pi[(\theta - \theta')^2 + 4]} \Rightarrow \| p_\theta - p_{\theta'} \|_{L^2}^2 = \frac{1}{\pi} \left(1 - \frac{1}{\frac{(\theta-\theta')^2}{4} + 1} \right).
$$

We now use the above remarks and check (7). Note that

$$
\mathcal{F}[p_\theta - p_{\theta'}](t) = [\exp(-it\theta) - \exp(-it\theta')] \exp(-|t|)
$$

and so

$$
\begin{aligned}
\frac{\mathbb{D}_{k_\gamma}^2(P_\theta, P_\theta')}{\| p_\theta - p_\theta' \|_{L^2}^2} &= \frac{\pi \int \gamma \mu(\gamma t) \left| [\exp(-it\theta) - \exp(-it\theta')] \exp(-|t|) \right|^2}{1 - \frac{1}{\frac{(\theta-\theta')^2}{4} + 1}} \\
&\geq \frac{\pi \int \gamma \mu(\gamma t) \left| [\exp(-it\theta) - \exp(-it\theta')] \exp(-t^2 - 1) \right|^2}{1 - \frac{1}{\frac{(\theta-\theta')^2}{4} + 1}}
\end{aligned}
$$

thanks to $|x| \leq x^2 + 1$. Identify the Fourier transform of the Gaussian density on the numerator to get

$$
\begin{aligned}
\frac{\mathbb{D}_{k_\gamma}^2(P_\theta, P_\theta')}{\| p_\theta - p_\theta' \|_{L^2}^2} &= \frac{2\pi \left[1 - \exp\left(-\frac{\|\theta - \theta'\|^2}{2(\gamma+2)} \right) \right]}{\exp(2) \sqrt{1 + \frac{2}{\gamma}} \left(1 - \frac{1}{\frac{(\theta-\theta')^2}{4} + 1} \right)} \\
&= \frac{2\pi \left[1 - \exp\left(-\frac{\|\theta - \theta'\|^2}{8} \right) \right]}{\exp(2) \sqrt{2} \left(1 - \frac{1}{\frac{(\theta-\theta')^2}{4} + 1} \right)}
\end{aligned}
$$

when $\gamma = 2$. For $x > 0$ we prove easily that

$$
\frac{1 - \exp(-x/2)}{1 - \frac{1}{1+x}} > \frac{1}{2},
$$

so

$$
\frac{\mathbb{D}_{k_\gamma}^2(P_\theta, P_\theta')}{\| p_\theta - p_\theta' \|_{L^2}^2} \geq \frac{\pi}{\sqrt{2} \exp(2)} \geq \frac{1}{4}.
$$

Now, in the adversarial contamination case,

$$
\begin{aligned}
(\hat{\theta}_n - \theta_0)^2 &= 4 \left[1 - \frac{1}{1 - \pi \| p_{\tilde{\theta}} - p_{\theta_0} \|^2} \right] \\
&\leq 4 \left[1 - \frac{1}{1 - 4\pi \mathbb{D}_{k_\gamma}^2(P_{\tilde{\theta}}, P_{\theta_0})} \right]
\end{aligned}
$$

$$\leq 4\left[1-\frac{1}{1-128\pi\left(\varepsilon^2+\frac{2+4\log(1/\delta)}{n}\right)}\right]$$

where the last inequality comes from Proposition 3.5. □

Remark A.1. *We applied the technique to the translated Cauchy model. Note that it is possible to apply it to many other translation models. However, in the translated uniform model (not reported in this paper), it leads to a suboptimal rate of convergence: $1/\sqrt{n}$, while the MLE is is $1/n$. But in the simulations we made, we observed that the MMD works very well in the uniform model – on the condition that γ is taken very small. We thus wonder whether it is possible, with another proof technique, to prove that the MMD estimator with $\gamma=\gamma(n)\to 0$ when $n\to\infty$ reaches the optimal rate of convergence. This is still an open question.*

Proof of Proposition 4.4. We remind that $P_{0:t}$ is the distribution of (X_0, X_t). Then

$$\begin{aligned}
\varrho_t &= \left|\mathbb{E}\left\langle \mu_{\delta_{X_t}}-\mu_{P^0},\mu_{\delta_{X_0}}-\mu_{P^0}\right\rangle_{\mathcal{H}_k}\right| \\
&= \left|\int k(x,y)P_{0:t}(\mathrm{d}(x,y)) - \iint k(x,y)P^0(\mathrm{d}x)P^0(\mathrm{d}y)\right| \\
&= \left|\int\left(\int \mathbf{1}_{\{u\geq\|x-y\|\}}f(u)\mathrm{d}u\right)P_{0:t}(\mathrm{d}(x,y)) - \iint\left(\int \mathbf{1}_{\{u>\|x-y\|\}}\right)f(u)\mathrm{d}uP^0(\mathrm{d}x)P^0(\mathrm{d}y)\right| \\
&= \left|\int_0^\infty\left(\int \mathbf{1}_{\{u\geq\|x-y\|\}}P_{0:t}(\mathrm{d}(x,y)) - \iint \mathbf{1}_{\{u\geq\|x-y\|\}}\ P^0(\mathrm{d}x)P^0(\mathrm{d}y)\right)f(u)\mathrm{d}u\right|.
\end{aligned}$$

For any partition $(A_i)_{i\in I}$ of $\mathbb{R}^d$ denote $I(u)=\{i\in I:(x,y)\in A_i^2\Rightarrow\|x-y\|\leq u\}$. Then

$$\sum_{i\in I(u)}\mathbf{1}_{A_i}(x)\mathbf{1}_{A_i}(y)\leq \mathbf{1}_{\{\|x-y\|\leq u\}}$$

and moreover $\mathbf{1}_{\{\|x-y\|\leq u\}}$ is the supremum of this sum over all possible measurable partitions, that is, for any $\varepsilon>0$, we can find a partition $(A_i)_{i\in I}$ such that

$$\mathbf{1}_{\{\|x-y\|\leq u\}}-\varepsilon\leq\sum_{i\in I(u)}\mathbf{1}_{A_i}(x)\mathbf{1}_{A_i}(y)\leq \mathbf{1}_{\{\|x-y\|\leq u\}}.$$

So,

$$\begin{aligned}
\varrho_t &\leq \left|\int_0^\infty\sum_{i\in I(u)}[P_{0:t}(A_i\times A_i)-P^0(A_i)^2]f(u)\mathrm{d}u\right|+\varepsilon \\
&\leq \int_0^\infty\sum_{i\in I(u)}\left|P_{0:t}(A_i\times A_i)-P^0(A_i)^2\right|f(u)\mathrm{d}u+\varepsilon \\
&\leq \int_0^\infty\sum_{i\in I}\left|P_{0:t}(A_i\times A_i)-P^0(A_i)^2\right|f(u)\mathrm{d}u+\varepsilon \\
&\leq \int_0^\infty 2\beta_t f(u)\mathrm{d}u+\varepsilon = 2\beta_t+\varepsilon.
\end{aligned}$$

□

Proof of Proposition 4.5. We start by a few preliminary remarks on (X_t). First, note that, for any ℓ and t,

$$X_{\ell+t}=\underbrace{\varepsilon_{\ell+t}+A\varepsilon_{\ell+t-1}+\cdots+A^{t-1}\varepsilon_{\ell+1}}_{=:B_\ell^t}+A^tX_\ell$$

where B_ℓ^t is independent of X_ℓ. Let P_b^t denote the distribution of B_ℓ^t (it does not depend on ℓ because of the stationarity).

Also, note that,

$$X_1 = AX_0 + \varepsilon_1$$

and thus

$$\|X_1\| \leq \|A\| \|X_0\| + \|\varepsilon_1\|. \tag{11}$$

Taking the expectation of (11), and using the stationarity, gives

$$\mathbb{E}(|X_0|) = \mathbb{E}(|X_1|) \leq \|A\| \mathbb{E}(\|X_0\|) + \mathbb{E}(\|\varepsilon_1\|)$$

which leads to

$$\mathbb{E}(\|X_0\|) \leq \frac{\mathbb{E}(\|\varepsilon_1\|)}{1-|a|} = \frac{\mathbb{E}(\|\varepsilon_0\|)}{1-\|a\|}.$$

In the same way, if $\|\varepsilon_t\| < c$ almost surely for any t, taking the supremum of (11) and using the stationarity gives, almost surely,

$$\|X_0\| \leq \frac{c}{1-\|a\|}.$$

We are now ready to start the derivation of the upper bound for ϱ_t. We have

$$\begin{aligned}
\varrho_t &= \left| \int k(x,y) P_{0:t}(\mathrm{d}(x,y)) - \iint k(x,y) P^0(\mathrm{d}x) P^0(\mathrm{d}y) \right| \\
&= \left| \iint k(x, A^t x + b) P^0(\mathrm{d}x) P_b^t(\mathrm{d}b) - \iiint k(x, A^t x' + b) P^0(\mathrm{d}x) P^0(\mathrm{d}x') P_b^t(\mathrm{d}b) \right| \\
&\leq \iiint \left| k(x, A^t x + b) - k(x, A^t x' + b) \right| P^0(\mathrm{d}x) P^0(\mathrm{d}x') P_b^t(\mathrm{d}b) \\
&\leq \iiint L \left| (A^t x + b) - (A^t x' + b) \right| P^0(\mathrm{d}x) P^0(\mathrm{d}x') P_b^t(\mathrm{d}b) \\
&= \iint L \|A\|^t \left| x - x' \right| P^0(\mathrm{d}x) P^0(\mathrm{d}x') \\
&\leq \int 2L \|A\|^t \left\| x \right\| P^0(\mathrm{d}x) \\
&= 2L \|A\|^t \mathbb{E}(\|X_0\|) \\
&\leq \frac{2L\|A\|^t \mathbb{E}(\|X_0\|)}{1-\|A\|}.
\end{aligned}$$

Let us now check Assumption 3.1. In order to do so, fix $\ell \in \{1, \ldots, n-1\}$ and a function $g : \mathcal{H}_k^\ell \to \mathbb{R}$ such that

$$|g(a_1, \ldots, a_\ell) - g(b_1, \ldots, b_\ell)| \leq \sum_{i=1}^{\ell} \|a_i - b_i\|_{\mathcal{H}_k}. \tag{12}$$

As we know that $\|\varepsilon_t\| \leq c$ a.s, we also have $\|X_t\| \leq c/(1-\|A\|)$. Fix $(x_1, \ldots, x_\ell)$ with $\|x_t\| \leq c/(1-\|A\|)$. We have

$$\begin{aligned}
&\left| \mathbb{E}[g(\mu_{\delta_{X_{\ell+1}}}, \ldots, \mu_{\delta_{X_n}}) | X_1 = x_1, \ldots, X_\ell = x_\ell] - \mathbb{E}[g(\mu_{\delta_{X_{\ell+1}}}, \ldots, \mu_{\delta_{X_n}})] \right| \\
&\quad = \left| \mathbb{E}[g(\mu_{\delta_{Ax_\ell + B_\ell^1}}, \ldots, \mu_{\delta_{A^{n-\ell-1} x_\ell + B_\ell^{n-\ell-1}}})] - \mathbb{E}[g(\mu_{\delta_{AX_\ell + B_\ell^1}}, \ldots, \mu_{\delta_{A^{n-\ell-1} X_\ell + B_\ell^{n-\ell-1}}})] \right| \\
&\quad \leq \mathbb{E}\left[\left| g(\mu_{\delta_{Ax_\ell + B_\ell^1}}, \ldots, \mu_{\delta_{A^{n-\ell-1} x_\ell + B_\ell^{n-\ell-1}}}) - \mathbb{E} g(\mu_{\delta_{AX_\ell + B_\ell^1}}, \ldots, \mu_{\delta_{A^{n-\ell-1} X_\ell + B_\ell^{n-\ell-1}}}) \right| \right]
\end{aligned}$$

$$\le \mathbb{E}\left[\sum_{i=1}^{n-\ell-1}\left\|\mu_{\delta_{A^i x_\ell+B_\ell^i}}-\mu_{\delta_{A^i X_\ell+B_\ell^i}}\right\|_{\mathcal{H}_k}\right] \text{ by (12)}$$

$$= \mathbb{E}\left[\sum_{i=1}^{n-\ell-1}\sqrt{k(A^i x_\ell+B_\ell^i, A^i x_\ell+B_\ell^i)+k(A^i X_\ell+B_\ell^i, A^i X_\ell+B_\ell^i)-2k(A^i x_\ell+B_\ell^i, A^i X_\ell+B_\ell^i)}\right].$$

We remind the assumption that $k(x,y)=F(|x-y|)$ where F is L-Lipschitz. So, in particular:

$$k(A^i x_\ell+B_\ell^i, A^i x_\ell+B_\ell^i)-k(A^i X_\ell+B_\ell^i, A^i x_\ell+B_\ell^i)\le L\|A\|^i\|x_\ell-X_\ell\|$$

and in the same way,

$$k(A^i X_\ell+B_\ell^i, A^i X_\ell+B_\ell^i)-k(A^i X_\ell+B_\ell^i, A^i x_\ell+B_\ell^i)\le L\|a\|^i\|x_\ell-X_\ell\|,$$

so:

$$\left|\mathbb{E}[g(\mu_{\delta_{X_{\ell+1}}},\dots,\mu_{\delta_{X_n}})|X_1=x_1,\dots,X_\ell=x_\ell]-\mathbb{E}[g(\mu_{\delta_{X_{\ell+1}}},\dots,\mu_{\delta_{X_n}})]\right|\le \mathbb{E}\left[\sum_{i=1}^{n-\ell}\sqrt{2L\|A\|^i\|x_\ell-X_\ell\|}\right].$$

We now use $\|X_t\|\le c/(1-\|A\|)$ a.s to get:

$$\left|\mathbb{E}[g(\mu_{\delta_{X_{\ell+1}}},\dots,\mu_{\delta_{X_n}})|X_1=x_1,\dots,X_\ell=x_\ell]-\mathbb{E}[g(\mu_{\delta_{X_{\ell+1}}},\dots,\mu_{\delta_{X_n}})]\right|\le \sum_{i=1}^{n-\ell-1}\frac{2Lc\|A\|^{\frac{i}{2}}}{1-\|A\|}=\sum_{i=1}^{n-\ell-1}\gamma_i.$$

This ends the proof. □

Appendix B: Proofs of Section 5

Proof of Proposition 5.1. Note that we can rewrite

$$\text{Crit}(\theta)=\iint k(x,x')p_\theta(x)p_\theta(x')\mu(\mathrm{d}x)\mu(\mathrm{d}x')-\frac{2}{n}\sum_{i=1}^n\int k(x,X_i)p_\theta(x)\mu(\mathrm{d}x).$$

The assumption of the proposition ensure that we can interchange the ∇ and $\iint$ symbols, and so

$$\begin{aligned}\nabla_\theta\text{Crit}(\theta)&=\iint k(x,x')\nabla_\theta[p_\theta(x)p_\theta(x')]\mu(\mathrm{d}x)\mu(\mathrm{d}x')-\frac{2}{n}\sum_{i=1}^n\int k(x,X_i)\nabla_\theta[p_\theta(x)]\mu(\mathrm{d}x)\\&=2\iint k(x,x')p_\theta(x)p_\theta(x')\nabla_\theta[\log p_\theta(x)]\mu(\mathrm{d}x)\mu(\mathrm{d}x')-\frac{2}{n}\sum_{i=1}^n\int k(x,X_i)\nabla_\theta[\log p_\theta(x)]p_\theta(x)\mu(\mathrm{d}x)\\&=2\mathbb{E}_{X,X'\sim P_\theta}\{k(X,X')\nabla_\theta[\log p_\theta(X)]\}-\frac{2}{n}\sum_{i=1}^n\mathbb{E}_{X\sim P_\theta}\{k(X_i,X)\nabla_\theta[\log p_\theta(X)]\}\\&=2\mathbb{E}_{X,X'\sim P_\theta}\left\{\left[k(X,X')-\frac{1}{n}\sum_{i=1}^n k(X_i,X)\right]\nabla_\theta[\log p_\theta(X)]\right\}.\end{aligned}$$

This ends the proof. □

Proof of Proposition 5.2. The assumption that Θ is bounded with radius $\mathcal{D}$ ensures that (2.17) in [Nemirovski et al., 2009] is satisfied, and the assumption on the expectation of the norm of the gradient ensures that (2.5) in [Nemirovski et al., 2009] is also satisfied. Thus, (2.21) is also satisfied, ant that is exactly the statement of (3) in Proposition 5.2. Then, we have:

$$\begin{aligned}\mathbb{E}\left[\mathbb{D}_k\left(P_{\hat\theta_n^{(T)}}, P^0\right)\right] &\leq \mathbb{D}_k\left(P_{\hat\theta_n^{(T)}}, \hat P_n\right) + \mathbb{D}_k\left(\hat P_n, P^0\right)\\ &= \sqrt{\mathbb{D}_k^2\left(P_{\hat\theta_n^{(T)}}, \hat P_n\right)} + \mathbb{D}_k\left(\hat P_n, P^0\right)\\ &\leq \sqrt{\mathbb{D}_k^2\left(P_{\hat\theta_n}, \hat P_n\right) + \frac{\mathcal{D}M}{\sqrt{T}}} + \mathbb{D}_k\left(\hat P_n, P^0\right)\end{aligned}$$

thanks to (3) in the proposition. We upper bound the second term thanks to Lemma 7.1 of the paper:

$$\mathbb{D}_k\left(\hat P_n, P^0\right) \leq \sqrt{\frac{1+2\sum_{t=1}^n \varrho_t}{n}}.$$

For the first term, we use:

$$\begin{aligned}\sqrt{\mathbb{D}_k^2\left(P_{\hat\theta_n}, \hat P_n\right) + \frac{\mathcal{D}M}{\sqrt{T}}} &\leq \mathbb{D}_k\left(P_{\hat\theta_n}, \hat P_n\right) + \sqrt{\frac{\mathcal{D}M}{\sqrt{T}}}\\ &\leq \inf_{\theta\in\Theta} \mathbb{D}_k\left(P_\theta, \hat P_n\right) + 2\sqrt{\frac{1+2\sum_{t=1}^n \varrho_t}{n}} + \sqrt{\frac{\mathcal{D}M}{\sqrt{T}}}\end{aligned}$$

thanks to Theorem 3.1 of the paper. Putting everything together leads to

$$\mathbb{E}\left[\mathbb{D}_k\left(P_{\hat\theta_n^{(T)}}, P^0\right)\right] \leq \inf_{\theta\in\Theta} \mathbb{D}_k\left(P_\theta, \hat P_n\right) + 3\sqrt{\frac{1+2\sum_{t=1}^n \varrho_t}{n}} + \sqrt{\frac{\mathcal{D}M}{\sqrt{T}}}$$

which ends the proof. □

References

[Chérief-Abdellatif and Alquier, 2021] Chérief-Abdellatif, B.-E. and Alquier, P. (2021). *Finite sample properties of parametric MMD estimation: robustness to misspecification and dependence.* Bernoulli, to appear.

[Nemirovski et al., 2009] Nemirovski, A., Juditsky, A., Lan, G., & Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4), 1574-1609.

[Sriperumbudur et al., 2010] Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., & Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures, *Journal of Machine Learning Research*, 11(Apr), 1517-1561.