A Proof of Proposition 4.2

By the identification in Proposition 4.1, we have $\gamma = \sum_{j=1}^k \frac{1}{\lambda_j} C_j^0 \otimes C_j^1$. We perform a bias-variance decomposition:

$$\begin{aligned}\int \|x-y\|^2 \,\mathrm{d}\gamma(x,y) &= \sum_{j=1}^k \frac{1}{\lambda_j} \int \|x-y\|^2 \,\mathrm{d}C_j^0(x)\mathrm{d}C_j^1(y) \\ &= \sum_{j=1}^k \frac{1}{\lambda_j} \int \|x-\mu(C_j^0) - (y-\mu(C_j^1)) + (\mu(C_j^0)-\mu(C_j^1))\|^2 \,\mathrm{d}C_j^0(x)\mathrm{d}C_j^1(y) \\ &= \sum_{j=1}^k \int \|x-\mu(C_j^0)\|^2 \,\mathrm{d}C_j^0(x) + \int \|y-\mu(C_j^1)\|^2 \mathrm{d}C_j^1(y) + \lambda_j \|\mu(C_j^0)-\mu(C_j^1)\|^2 ,\end{aligned}$$

where the cross terms vanish by the definition of $\mu(C_j^0)$ and $\mu(C_j^1)$. □

B Proof of Proposition 4.3

We first show that if H is an optimal solution to (5), then the hubs $z_1, \ldots, z_k$ satisfy $z_j = \frac{1}{2}(\mu(C_j^0) + \mu(C_j^1))$ for $j = 1, \ldots k$. Let P be any distribution in $\mathcal{D}_k$. Denote the support of P by $z_1, \ldots, z_k$, and let $\{C_j^0\}, \{C_j^1\}$ be the partition of $\hat{P}_0$ and $\hat{P}_1$ induced by the objective $W_2^2(P, \hat{P}_0) + W_2^2(P, \hat{P}_1)$. By the same bias-variance decomposition as in the proof of Proposition 4.2,

$$W_2^2(\hat{P}_0, P) = \sum_{j=1}^k \int_{C_j^0} \|x - z_j\|^2 \,\mathrm{d}\hat{P}_0(x) = \sum_{j=1}^k \int_{C_j^0} \|x-\mu(C_j^0)\|^2 \,\mathrm{d}\hat{P}_0(x) + \lambda_j \|z_j - \mu(C_j^0)\|^2 ,$$

and since the analogous claim holds for $\hat{P}_1$, we obtain that

$$W_2^2(P,\hat{P}_0)+W_2^2(P,\hat{P}_1) = \sum_{j=1}^k \int_{C_j^0} \|x-\mu(C_j^0)\|^2 \,\mathrm{d}\hat{P}_0(x) + \int_{C_j^1} \|y-\mu(C_j^1)\|^2 \,\mathrm{d}\hat{P}_1(y) + \lambda_j(\|z_j-\mu(C_j^0)\|^2 + \|z_j-\mu(C_j^1)\|^2) .$$

The first two terms depend only on the partitions of $\hat{P}_0$ and $\hat{P}_1$, and examining the final term shows that any minimizer of $W_2^2(P, \hat{P}_0) + W_2^2(P, \hat{P}_1)$ must have $z_j = \frac{1}{2}(\mu(C_j^0) + \mu(C_j^1))$ for $j = 1, \ldots k$, where C_j^0 and C_j^1 are induced by P, in which case $\|z_j - \mu(C_j^0)\|^2 + \|z_j - \mu(C_j^1)\|^2 = \frac{1}{2}\|\mu(C_j^0) - \mu(C_j^1)\|^2$. Minimizing over $P \in \mathcal{D}_k$ yields the claim. □

C Proof of Theorem 4

The proof of Theorem 4 relies on the following propositions, which shows that controlling the gap between $W_2^2(\rho, P)$ and $W_2^2(\rho, Q)$ is equivalent to controlling the distance between P and Q with respect to a simple integral probability metric [Müller, 1997].

We make the following definition.

Definition 5. *A set $S \in \mathbb{R}^d$ is a* n-polyhedron *if S can be written as the intersection of n closed half-spaces.*

We denote the set of n-polyhedra by $\mathcal{P}_n$. Given $c \in \mathbb{R}^d$ and $S \in \mathcal{P}_{k-1}$, define

$$f_{c,S}(x) := \|x - c\|^2 \mathbb{1}_{x \in S} \qquad \forall x \in \mathbb{R}^d .$$

Proposition C.1. *Let P and Q be probability measures supported on the unit ball in $\mathbb{R}^d$. The*

$$\sup_{\rho \in \mathcal{D}_k} |W_2^2(\rho, P) - W_2^2(\rho, Q)| \le 5k \sup_{c:\|c\| \le 1, S \in \mathcal{P}_{k-1}} |\mathbb{E}_P f_{c,S} - \mathbb{E}_Q f_{c,S}| . \tag{7}$$

To obtain Theorem 4, we use techniques from empirical process theory to control the right side of (7) when $Q = \hat{P}$.

Proposition C.2. *There exists a universal constant C such that, if P is supported on the unit ball and $X_1, \dots, X_n \sim \mu$ are i.i.d., then*

$$\mathbb{E} \sup_{c:\|c\|\le 1, S\in\mathcal{P}_{k-1}} |\mathbb{E}_P f_{c,S} - \mathbb{E}_{\hat{P}} f_{c,S}| \le C\sqrt{\frac{kd\log k}{n}}\,.$$

With these tools in hand, the proof of Theorem 4 is elementary.

Proof of Theorem 4. Proposition C.1 implies that

$$\mathbb{E} \sup_{\rho\in\mathcal{D}_k} |W_2^2(\rho,\hat{\mu}) - W_2^2(\rho,\mu)| \lesssim \sqrt{\frac{k^3 d\log k}{n}}\,.$$

To show the high probability bound, it suffices to apply the bounded difference inequality (see [McDiarmid, 1989]) and note that, if $\hat{P}$ and $\tilde{P}$ differ in the location of a single sample, then for any ρ, we have the bound $|W_2^2(\rho,\hat{P}) - W_2^2(\rho,\tilde{P})| \le 4/n$. The concentration inequality immediately follows. □

We now turn to the proofs of Propositions C.1 and Propositions C.2.

We first review some facts from the literature. It is by now well known that there is an intimate connection between the k-means objective and the squared Wasserstein 2-distance [Canas and Rosasco, 2012, Ng, 2000, Pollard, 1982]. This correspondence is based on the following observation, more details about which can be found in [Graf and Luschgy, 2000]: given fixed points $c_1, \dots, c_k$ and a measure P, consider the quantity

$$\min_{w\in\Delta_k} W_2^2\Big(\sum_{i=1}^k w_i\delta_{c_i}, P\Big)\,, \tag{8}$$

where the minimization is taken over all probability vectors $w := (w_1, \dots, w_k)$. Note that, for *any* measure ρ supported on $\{c_1, \dots, c_k\}$, we have the bound

$$W_2^2(\rho,P) \ge \mathbb{E}\big[\min_{i\in[k]} \|X - c_k\|^2\big] \qquad X\sim P\,.$$

On the other hand, this minimum can be achieved by the following construction. Denote by $\{S_1, \dots, S_k\}$ the Voronoi partition [Okabe et al., 2000] of $\mathbb{R}^d$ with respect to the centers $\{c_1, \dots, c_k\}$ and let $\rho = \sum_{i=1}^k P(S_i)\delta_{c_i}$. If we let $T : \mathbb{R}^d \to \{c_1, \dots, c_k\}$ be the function defined by $S_i = T^{-1}(c_i)$ for $i \in [k]$, then $(\mathrm{id}, T)_\sharp P$ defines a coupling between P and ρ which achieves the above minimum, and

$$\mathbb{E}[\|X - T(X)\|^2] = \mathbb{E}[\min_{i\in[k]} \|X - c_i\|^2] \qquad X\sim P\,.$$

The above argument establishes that the measure closest to P with prescribed support of at most k points is induced by a Voronoi partition of $\mathbb{R}^d$, and this observation carries over into the context of the k-means problem [Canas and Rosasco, 2012], where one seeks to solve

$$\min_{\rho\in\mathcal{D}_k} W_2^2(\rho,P)\,. \tag{9}$$

The above considerations imply that the minimizing measure will correspond to a Voronoi partition, and that the centers $c_1, \dots, c_k$ will lie at the centroids of each set in the partition with respect to P. As above, there will exist a map T realizing the optimal coupling between P and ρ, where the sets $T^{-1}(c_i)$ for $i \in [k]$ form a Voronoi partition of $\mathbb{R}^d$. In particular, standard facts about Voronoi cells for the ℓ_2 distance [Okabe et al., 2000, Definition V4] imply that, for $i \in [k]$, the set $\mathrm{cl}(T^{-1}(c_i))$ is a $(k-1)$-polyhedron. (See Definition 5 above.)

In the case when ρ is an *arbitrary* measure with support of size k—and not the solution to an optimization problem such as (8) or (9)—it is no longer the case that the optimal coupling between P and ρ corresponds to a Voronoi partition of $\mathbb{R}^d$. The remainder of this section establishes, however, that, if P is absolutely continuous with respect to the Lebgesgue measure, then there does exist a map T such that the fibers of points in the image of T have a particularly simple form: like Voronoi cells, the sets $\{\mathrm{cl}(T^{-1}(c_i)\}_{i=1}^k$ can be taken to be simple polyehdra.

Definition 6. *A function $T : \mathbb{R}^d \to \mathbb{R}^d$ is a* polyhedral quantizer *of order k if T takes at most k values and if, for each $x \in \mathrm{Im}(T)$, the set $\mathrm{cl}(T^{-1}(x))$ is a $(k-1)$-polyhedron and $\partial T^{-1}(x)$ has zero Lebesgue measure.*

We denote by $\mathcal{Q}_k$ the set of k-polyhedral quantizers whose image lies inside the unit ball of $\mathbb{R}^d$.

Proposition C.3. *Let P be any absolutely continuous measure in $\mathbb{R}^d$, and let ρ be any measure supported on k points. Then there exists a map T such that $(id, T)_\sharp P$ is an optimal coupling between P and ρ and T is a polyhedral quantizer of order k.*

Proof. Denote by $\rho_1, \ldots, \rho_k$ the support of ρ. Standard results in optimal transport theory [Santambrogio, 2015, Theorem 1.22] imply that there exists a convex function u such that the optimal coupling between P and ρ is of the form $(\mathrm{id}, \nabla u)_\sharp P$. Let $S_i = (\nabla u)^{-1}(\rho_i)$.

Since $\nabla u(x) = \rho_j$ for any $x \in S_j$, the restriction of u to S_j must be an affine function. We obtain that there exists a constant β_j such that

$$u(x) = \langle \rho_j, x \rangle + \beta_j \qquad \forall x \in S_j \,.$$

Since ρ_j has nonzero mass, the fact that $\nabla u_\sharp P = \rho$ implies that $P(S_j) > 0$, and, since P is absolutely continuous with respect to the Lebesgue measure, this implies that S_j has nonempty interior. If $x \in \mathrm{int}(S_j)$, then $\partial u(x) = \{\rho_j\}$. Equivalently, for all $y \in \mathbb{R}^d$,

$$u(y) \geq \langle \rho_j, y \rangle + \beta_j \,.$$

Employing the same argument for all $j \in [k]$ yields

$$u(x) \geq \max_{j \in [k]} \langle \rho_j, x \rangle + \beta_j \,.$$

On the other hand, if $x \in S_i$, then

$$u(x) = \langle \rho_i, x \rangle + \beta_i \leq \max_{j \in [k]} \langle \rho_j, x \rangle + \beta_j \,.$$

We can therefore take u to be the convex function

$$u(x) = \max_{j \in [k]} \langle \rho_j, x \rangle + \beta_j \,,$$

which implies that, for $i \in [k]$,

$$\begin{aligned} \mathrm{cl}(S_i) &= \{y \in \mathbb{R}^d : \langle \rho_i, x \rangle + \beta_i \geq \langle \rho_j, x \rangle + \beta_j \quad \forall j \in [k] \setminus \{i\}\} \\ &= \bigcap_{j \neq i} \{y \in \mathbb{R}^d : \langle \rho_i, x \rangle + \beta_i \geq \langle \rho_j, x \rangle + \beta_j\} \,. \end{aligned}$$

Therefore $\mathrm{cl}(S_i)$ can be written as the intersection of $k-1$ halfspaces. Moreover, $\partial S_i \subseteq \bigcup_{j \neq i} \{y \in \mathbb{R}^d : \langle \rho_i, x \rangle + \beta_i = \langle \rho_j, x \rangle + \beta_j\}$, which has zero Lebesgue measure, as claimed. □

C.1 Proof of Proposition C.1

By symmetry, it suffices to show the one-sided bound

$$\sup_{\rho \in \mathcal{D}_k} W_2^2(\rho, Q) - W_2^2(\rho, P) \leq 5k \sup_{c : \|c\| \leq 1, S \in \mathcal{P}_{k-1}} |\mathbb{E}_P f_{c,S} - \mathbb{E}_Q f_{c,S}| \,.$$

We first show the claim for P and Q which are absolutely continuous. Fix a $\rho \in \mathcal{D}_k$. Since P and Q are absolutely continuous, we can apply Proposition C.3 to obtain that there exists a $T \in \mathcal{Q}_k$ such that

$$W_2^2(\rho, P) = \mathbb{E}_P \|X - T(X)\|^2 \,.$$

Let $\{c_1, \ldots, c_k\}$ be the image of T, and for $i \in [k]$ let $S_i := \mathrm{cl}(T^{-1}(c_i))$. Denote by $d_{TV}(\mu, \nu) := \sup_{A \text{ measurable}} |\mu(A) - \nu(A)|$ the total variation distance between μ and ν. Applying Lemma E.1 to ρ and Q yields that

$$W_2^2(Q, \rho) \leq \mathbb{E}_Q \|X - T(X)\|^2 + 4 \mathrm{d_{TV}}(T_\sharp Q, \rho) \,.$$

Since $\rho = T_\sharp P$ and Q and P are absolutely continuous with respect to the Lebesgue measure, we have

$$\begin{aligned}\mathrm{d_{TV}}(T_\sharp Q, \rho) &= \mathrm{d_{TV}}(T_\sharp Q, T_\sharp P)\\ &= \frac{1}{2}\sum_{i=1}^{k} |P(T^{-1}(c_i)) - Q(T^{-1}(c_i))|\\ &= \frac{1}{2}\sum_{i=1}^{k} |P(S_i) - Q(S_i)|\,.\end{aligned}$$

Combining the above bounds yields

$$\begin{aligned}W_2^2(\rho, Q) - W_2^2(\rho, P) &\le \mathbb{E}_Q\|X - T(X)\|^2 - \mathbb{E}_P\|X - T(X)\|^2 + 2\sum_{i=1}^{k}|P(S_i) - Q(S_i)|\\ &\le \sum_{i=1}^{k} |\mathbb{E}_Q\|X - c_i\|^2 \mathbb{1}_{X\in S_i} - \mathbb{E}_P\|X - c_i\|^2\mathbb{1}_{X\in S_i}| + 2|P(S_i) - Q(S_i)|\\ &\le k\sup_{c,S}\left(|\mathbb{E}_Q\|X - c\|^2\mathbb{1}_{X\in S} - \mathbb{E}_P\|X - c\|^2\mathbb{1}_{X\in S}| + 2|P(S) - Q(S)|\right)\\ &= k\sup_{c,S}\left(|\mathbb{E}_Q f_{c,S} - \mathbb{E}_P f_{c,S}| + 2|P(S) - Q(S)|\right)\end{aligned}$$

where the supremum is taken over $c \in \mathbb{R}^d$ satisfying $\|c\| \le 1$ and $S \in \mathcal{P}_{k-1}$.

If $\|v\| = 1$, then

$$\mathbb{1}_{X\in S} = \frac{1}{2}(\|X + v\|^2 + \|X - v\|^2 - 2\|X\|^2)\mathbb{1}_{X\in S}\,,$$

which implies

$$\begin{aligned}|P(S) - Q(S)| &= |\mathbb{E}_P\mathbb{1}_{X\in S} - \mathbb{E}_Q\mathbb{1}_{X\in S}|\\ &\le \frac{1}{2}\left(|\mathbb{E}_P f_{v,S} - \mathbb{E}_Q f_{v,S}| + |\mathbb{E}_P f_{-v,S} - \mathbb{E}_Q f_{-v,S}| + 2|\mathbb{E}_P f_{0,S} - \mathbb{E}_Q f_{0,S}|\right)\\ &\le 2\sup_{c,S}|\mathbb{E}_P f_{c,S} - \mathbb{E}_Q f_{c,S}|\,.\end{aligned}$$

Combining the above bounds yields

$$W_2^2(\rho, Q) - W_2^2(\rho, P) \le 5k\sup_{c,S}|\mathbb{E}_P f_{c,S} - \mathbb{E}_Q f_{c,S}|$$

Finally, since this bound holds for all $\rho \in \mathcal{D}_k$, taking the supremum of the left side yields the claim for absolutely continuous P and Q.

To prove the claim for arbitrary measures, we reduce to the absolutely continuous case. Let $\delta \in (0, 1)$ be arbitrary, and let $\mathcal{K}_\delta$ be any absolutely continuous probability measure such that, if $Z \sim \mathcal{K}_\delta$ then $\|Z\| \le \delta$ almost surely. Let $\rho \in \mathcal{D}_k$. The triangle inequality for W_2 implies

$$|W_2(\rho, Q) - W_2(\rho, Q * \mathcal{K}_\delta)| \le W_2(Q, Q * \mathcal{K}_\delta) \le \delta\,,$$

where the final inequality follows from the fact that, if $X \sim Q$ and $Z \sim \mathcal{K}_\delta$, then $W_2^2(Q, Q * \mathcal{K}_\delta) \le \mathbb{E}\|X - (X + Z)\|^2 \le \delta^2$. Since ρ and Q are both supported on the unit ball, the trivial bound $W_2(\rho, Q) \le 2$ holds. If $\delta \le 1$, then $W_2(\rho, Q * \mathcal{K}_\delta) \le 3$, and we obtain

$$|W_2^2(\rho, Q) - W_2^2(\rho, Q * \mathcal{K}_\delta)| \le 5\delta\,.$$

The same argument implies

$$|W_2^2(\rho, P) - W_2^2(\rho, P * \mathcal{K}_\delta)| \le 5\delta\,.$$

Therefore

$$\sup_{\rho\in\mathcal{D}_K} W_2^2(\rho, Q) - W_2^2(\rho, P) \le \sup_{\rho\in\mathcal{D}_K} W_2^2(\rho, Q * \mathcal{K}_\delta) - W_2^2(\rho, P * \mathcal{K}_\delta) + 10\delta\,.$$

Likewise, for any x and c in the unit ball, if $\|z\| \le \delta$, then by the exact same argument as was used above to bound $|W_2^2(\rho, Q) - W_2^2(\rho, Q * \mathcal{K}_\delta)|$, we have

$$|f_{c,S}(x+z) - f_{c,S-z}(x)| \le 5\delta\,.$$

Let $Z \sim \mathcal{K}_\delta$ be independent of all other random variables, and denote by $\mathbb{E}_Z$ expectation with respect to this quantity. Now, applying the proposition to the absolutely continuous measures $P * \mathcal{K}_\delta$ and $Q * \mathcal{K}_\delta$, we obtain

$$\begin{aligned}
\sup_{\rho \in \mathcal{D}_k} W_2^2(\rho, Q) - W_2^2(\rho, P) &\le 5k \sup_{c,S} |\mathbb{E}_Z\left[\mathbb{E}_P f_{c,S}(X+Z) - \mathbb{E}_Q f_{c,S}(X+Z)\right]| + 10\delta \\
&\le \mathbb{E}_Z\left[5k \sup_{c,S} |\mathbb{E}_P f_{c,S}(X+Z) - \mathbb{E}_Q f_{c,S}(X+Z)|\right] + 10\delta \\
&\le \mathbb{E}_Z\left[5k \sup_{c,S} |\mathbb{E}_P f_{c,S-Z} - \mathbb{E}_Q f_{c,S-Z}|\right] + 20\delta\,.
\end{aligned}$$

It now suffices to note that, for any $S \in \mathcal{P}_{k-1}$ and any $z \in \mathbb{R}^d$, the set $S - z \in \mathcal{P}_{k-1}$. In particular, this implies that

$$z \mapsto \sup_{c,S} |\mathbb{E}_P f_{c,S-z} - \mathbb{E}_Q f_{c,S-z}|$$

is constant, so that the expectation with respect to Z can be dropped.

We have shown that, for any $\delta \in (0,1)$, the bound

$$\sup_{\rho \in \mathcal{D}_K} W_2^2(\rho, P) - W_2^2(\rho, Q) \le 5k \sup_{c,S} |\mathbb{E}_P f_{c,S} - \mathbb{E}_Q f_{c,S}| + 20\delta$$

holds. Taking the infimum over $\delta > 0$ yields the claim. □

D Proof of Proposition C.2

In this proof, the symbol C will stand for a universal constant whose value may change from line to line. For convenience, we will use the notation $\sup_{c,S}$ to denote the supremum over the feasible set $c : \|c\| \le 1, S \in \mathcal{P}_{k-1}$.

We employ the method of [Maurer and Pontil, 2010]. By a standard symmetrization argument [Giné and Nickl, 2016], if $g_1, \dots, g_n$ are i.i.d. standard Gaussian random variables, then the quantity in question is bounded from above by

$$\frac{\sqrt{2\pi}}{n} \mathbb{E} \sup_{c,S} \left| \sum_{i=1}^n g_i f_{c,S}(X_i) \right| \le \frac{\sqrt{8\pi}}{n} \mathbb{E} \sup_{c,S} \sum_{i=1}^n g_i f_{c,S}(X_i) + \frac{C}{\sqrt{n}}\,.$$

Given c and c'' in the unit ball and $S, S' \in \mathcal{P}_{k-1}$, consider the increment $(f_{c,S}(x) - f_{c',S'}(x))^2$. If $x \in S \triangle S'$ and $\|x\| \le 1$, then

$$(f_{c,S}(x) - f_{c',S'}(x))^2 \le \max\left\{\|x-c\|^4, \|x-c'\|^4\right\} \le 16\,.$$

On the other hand, if $x \notin S \triangle S'$, then

$$(f_{c,S}(x) - f_{c',S'}(x))^2 \le (\|x-c\|^2 - \|x-c'\|^2)^2\,.$$

Therefore, for any x in the unit ball,

$$(f_{c,S}(x) - f_{c',S'}(x))^2 \le 16(\mathbb{1}_{x \in S} - \mathbb{1}_{x \in S'})^2 + (\|x-c\|^2 - \|x-c'\|^2)^2\,.$$

This fact implies that the Gaussian processes

$$\begin{aligned}
G_{c,S} &:= \sum_{i=1}^n g_i f_{c,S}(X_i) && g_i \sim \mathcal{N}(0,1) \text{ i.i.d} \\
H_{c,S} &:= \sum_{i=1}^n 4 g_i \mathbb{1}_{X_i \in S} + g_i' \|X_i - c\|^2 && g_i, g_i' \sim \mathcal{N}(0,1) \text{ i.i.d}\,,
\end{aligned}$$

satisfy

$$\mathbb{E}(G_{c,S} - G_{c',S'})^2 \le \mathbb{E}(H_{c,S} - H_{c',S'})^2 \qquad \forall c, c', S, S' .$$

Therefore, by the Slepian-Sudakov-Fernique inequality [Fernique, 1975, Slepian, 1962, Sudakov, 1971],

$$\begin{aligned}\mathbb{E}\sup_{c,S}\sum_{i=1}^n g_i f_{c,S}(X_i) &\le \mathbb{E}\sup_{c,S}\sum_{i=1}^n 4g_i \mathbb{1}_{X_i\in S} + g_i'\|X_i - c\|^2 \\ &\le \mathbb{E}\sup_{S\in\mathcal{P}_{k-1}} 4\sum_{i=1}^n g_i\mathbb{1}_{X_i\in S} + \mathbb{E}\sup_{c:\|c\|\le 1}\sum_{i=1}^n g_i\|X_i - c\|^2 .\end{aligned}$$

We control the two terms separately. The first term can be controlled using the VC dimension of the class $\mathcal{P}_{k-1}$ [Vapnik and Červonenkis, 1971] by a standard argument in empirical process theory (see, e.g., [Giné and Nickl, 2016]). Indeed, using the bound [Dudley, 1978, Lemma 7.13] combined with the chaining technique [Vershynin, 2016] yields

$$\mathbb{E}\sup_{S\in\mathcal{P}_{k-1}} 4\sum_{i=1}^n g_i\mathbb{1}_{X_i\in S} \le C\sqrt{n\mathrm{VC}(\mathcal{P}_{k-1})} .$$

By Lemma E.2, $VC(\mathcal{P}_{k-1}) \le Cdk\log k$; hence

$$\mathbb{E}\sup_{S\in\mathcal{P}_{k-1}} 4\sum_{i=1}^n g_i\mathbb{1}_{X_i\in S} \le C\sqrt{ndk\log k} .$$

The second term can be controlled as in [Maurer and Pontil, 2010, Lemma 3]:

$$\begin{aligned}\mathbb{E}\sup_{c:\|c\|\le 1}\sum_{i=1}^n g_i\|X_i - c\|^2 &= \mathbb{E}\sup_{c:\|c\|\le 1}\sum_{i=1}^n g_i(\|X_i\|^2 - 2\langle X_i, c\rangle + \|c\|^2) \\ &\le 2\mathbb{E}\sup_{c:\|c\|\le 1}\sum_{i=1}^n g_i\langle X_i, c\rangle + \sup_{c:\|c\|\le 1}\sum_{i=1}^n g_i\|c\|^2 \\ &\le 2\mathbb{E}\left\|\sum_{i=1}^n g_i X_i\right\| + \left|\sum_{i=1}^n g_i\right| \\ &\le C\sqrt{n}\end{aligned}$$

for some absolute constant C.

Combining the above bounds yields

$$\frac{\sqrt{8\pi}}{n}\mathbb{E}\sup_{c:\|c\|\le 1, S\in\mathcal{P}_{k-1}}\sum_{i=1}^n g_i f_{c,S}(X_i) \le C\sqrt{\frac{dk\log k}{n}} ,$$

and the claim follows. □

E Additional lemmas

Lemma E.1. *Let μ and ν be probability measures on $\mathbb{R}^d$ supported on the unit ball. If $T : \mathbb{R}^d \to \mathbb{R}^d$, then*

$$W_2^2(\mu,\nu) \le \mathbb{E}\|X - T(X)\|^2 + 4\mathrm{d_{TV}}(T_\sharp\mu, \nu) \qquad X\sim\mu .$$

Proof. If $X\sim\mu$, then $(X, T(X))$ is a coupling between μ and $T_\sharp\mu$. Combining this coupling with the optimal coupling between $T\sharp\mu$ and ν and applying the gluing lemma [Villani, 2009] yields that there exists a triple $(X, T(X), Y)$ such that $X\sim\mu$, $Y\sim\nu$, and $\mathbb{P}[T(X)\ne Y] = \mathrm{d_{TV}}(T_\sharp\mu,\nu)$.

$$\begin{aligned}W_2^2(\mu,\nu) &\le \mathbb{E}[\|X - Y\|^2] \\ &= \mathbb{E}[\|X - Y\|^2\mathbb{1}_{T(X)=Y}] + \mathbb{E}[\|X - Y\|^2\mathbb{1}_{T(X)\ne Y}] \\ &\le \mathbb{E}[\|X - T(X)\|^2] + 4\mathrm{d_{TV}}(T_\sharp\mu,\nu) ,\end{aligned}$$

where the last inequality uses the fact that $\mathbb{P}[T(X)\ne Y] = \mathrm{d_{TV}}(T_\sharp\mu,\nu)$ and that $\|X - Y\| \le 2$ almost surely. □

Lemma E.2. *The class* $\mathcal{P}_{k-1}$ *satisfies* $\mathrm{VC}(\mathcal{P}_{k-1}) \leq Cdk \log k$.

Proof. The claim follows from two standard results in VC theory:

- The class all half-spaces in dimension d has VC dimension $d+1$ [Devroye et al., 1996, Corollary 13.1].
- If $\mathcal{C}$ has VC dimension at most n, then the class $\mathcal{C}_s := \{c_1 \cap \ldots c_s : c_i \in \mathcal{C}\, \forall i \in [s]\}$ has VC dimension at most $2ns \log(3s)$ [Blumer et al., 1989, Lemma 3.2.3].

Since $\mathcal{P}_{k-1}$ consists of intersections of at most $k-1$ half-spaces, we have

$$\mathrm{VC}(\mathcal{P}_{k-1}) \leq 3(d+1)(k-1)\log(3(k-1)) \leq Cdk \log k$$

for a universal constant C. □

F Details on numerical experiments

In this section we present implementation details for our numerical experiments.

In all experiments, the relative tolerance of the objective value is used as a stopping criterion for FactoredOT. We terminate calculation when this value reached 10^{-6}.

F.1 Synthetic experiments from Section 6.1

In the synthetic experiments, the entropy parameter was set to 0.1.

F.2 Single cell RNA-seq batch correction experiments from Section 6.2

We obtained a pair of single cell RNA-seq data sets from Haghverdi et al. [2018]. The first dataset [Nestorowa et al., 2016] was generated using SMART-seq2 protocol [Picelli et al., 2014], while the second dataset [Paul et al., 2015] was generated using the MARS-seq protocol [Jaitin et al., 2014].

We preprocessed the data using the procedure described by Haghverdi et al. [2018] to reduce to 3,491 dimensions.

Nex, we run our domain adaptation procedure. To determine the choice of parameters, we perform cross-validation over 20 random sub-samples of the data, each containing 100 random cells of each of the three cell types in both source and target distribution. Performance is then determined by the mis-classification over 20 independent versions of the same kind of random sub-samples.

For all methods involving entropic regularization (FOT, OT-ER, OT-L1L2), the candidates for the entropy parameter are $\{10^{-3}, 10^{-2.5}, 10^{-2}, 10^{-1.5}, 10^{-1}\}$.

For FOT and k-means OT, the number of clusters is in $\{3, 6, 9, 12, 20, 30\}$.

For OT-L1L2, the regularization parameter is in $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$.

For all subspace methods (SA, TCA), the dimensionality is in $\{10, 20, \ldots, 70\}$.

The labels are determined by first adjusting the sample and then performing a majority vote among 20 nearest neighbors. While similar experiments [Courty et al., 2014, 2017, Pan et al., 2011] employed 1NN classification because it does not require a tuning parameter, we observed highly decreased performance among all considered domain adaptation methods and therefore chose to use a slightly stronger predictor. The results are not sensitive to the choice of k for the kNN predictor for $k \approx 20$.