A Component family and practical aspects of optimization

In order to use Algorithm 1, we need to compute or estimate $\langle h, \phi \rangle$ for any $\phi \in L^2(\mu)$ and $\langle h, h' \rangle$ for any $h, h' \in \mathcal{H}$. For arbitrary ϕ, we use Monte Carlo estimates based on samples from h^2 via

$$\forall \phi \in L^2(\mu), \qquad \langle h, \phi \rangle = \int h^2(x) \frac{\phi(x)}{h(x)} \mu(\mathrm{d}x) = \mathbb{E}_{h^2}\left[\frac{\phi(X)}{h(X)}\right] \approx \frac{1}{S}\sum_{s=1}^{S} \frac{\phi(X_s)}{h(X_s)} \quad X_s \overset{\text{i.i.d.}}{\sim} h^2,$$

and employ an exponential component family $\mathcal{H}$ such that inner products $\langle h, h' \rangle$ between members of $\mathcal{H}$ are available in closed-form. In other words, for some base density $k(x)$, sufficient statistic $T(x)$, and log-partition $A(\eta)$, we let

$$\mathcal{H} = \left\{ h_\eta \in L^2(\mu) : h_\eta^2(x) = k(x) \exp\left(\eta^T T(x) - A(\eta)\right) \right\}.$$

Denoting η_i to be the natural parameter for g_i, then $g_i = h_{\eta_i}$ and

$$Z_{ij} = \langle g_i, g_j \rangle = \int h_{\eta_i}(x) h_{\eta_j}(x) \mu(\mathrm{d}x) = \exp\left(A\left(\frac{\eta_i + \eta_j}{2}\right) - \frac{A(\eta_i) + A(\eta_j)}{2}\right). \tag{10}$$

In practice, we use a few techniques to improve the stability and performance of UBVI:

Component Initialization The performance of variational boosting methods is often sensitive to the choice of initialization in each component optimization. The initialization used in this work is based on the intuition that after the first component optimization, each subsequent optimization will typically do one of two things: either it will find a new mode, or it will attempt to refine a previously found mode. If we wish to refine a previous mode, it is useful to initialize the optimization near that mode with a similar covariance structure. If we wish to discover a new mode, it is preferable to sample an initialization from the present distribution with significant added noise. In the experimental section of this work, we take the middle ground. We first sample a component from the current mixture approximation. Then, we generate an initialization for the Gaussian mean by sampling from that component with its covariance increased by a factor of 16. Finally, we initialize the covariance by using that component's covariance multiplied by a standard log-normal random variable.

Objective Transformation We maximize $\log(J(x))\mathbb{1}\left[J(x) \geq 0\right] - \log(-J(x))\mathbb{1}\left[J(x) < 0\right]$, where $J(x)$ is the objective in Eq. (5), to avoid vanishing gradients and handle possible negativity; while this technically makes the Monte Carlo-based stochastic gradient estimates biased, it significantly improves performance in practice.

Parametrization The choice of parametrization can have a significant effect on the conditioning of the optimization problem. Although we exploit the properties of the exponential family for Z_{ij} evaluation in Eq. (10), we do not use the natural parametrization during optimization. In particular, we optimize over the mean and log-transformed marginal variances $\log \sigma_i^2$ in the diagonal covariance matrix $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_D^2)$.

Large-Scale Data If the target density p arises from a Bayesian posterior inference problem with a large dataset, computing p and its gradients exactly in each component optimization iteration is expensive. Thus, one can use a Monte Carlo minibatch approximation with uniformly subsampled data per [50].

Estimating $\langle f, g \rangle$ We use different numbers of samples for the component optimization stochastic gradient estimates and the estimates of $\langle f, g_n \rangle$ (Line 9, Algorithm 1) required to solve the UBVI weight optimization. In particular, we use a relatively high number of samples (10,000 in our experiments) for estimating $\langle f, g_n \rangle$, as these each need to be estimated only once, and they have a high impact on the choice of weights and thus future components; and for stochastic optimization, we use a lower number of samples (1,000 in our experiments) to avoid overly expensive component optimizations.

B Proofs

B.1 Proof of gradient boosting BVI behaviour

Proof of Proposition 1. Let $\phi(x; \sigma^2)$ be the normal density with mean 0 and variance σ^2. Then Eq. (1) is

$$\begin{aligned}
\sigma^{\star 2} &= \arg\min_{\sigma^2} \int \phi(x; \sigma^2) \log \frac{\phi(x; \sigma^2)^{r_2} \phi(x; \tau^2)}{\phi(x; 1)} \mathrm{d}x \\
&= \arg\min_{\sigma^2} -r_2 \log \sigma - \frac{\sigma^2}{2\tau^2} + \frac{\sigma^2}{2}
\end{aligned}$$

$$= \begin{cases} \infty & \tau^2 \leq 1 \\ \frac{r_2\tau^2}{\tau^2-1} & \tau^2 > 1. \end{cases}$$

Therefore, if the initialization has variance $\tau^2 \leq 1$ the component optimization is degenerate. Note that for any two variances σ_1^2, σ_2^2, the weight optimization is

$$w^\star = \arg\min_{w\in[0,1]} \int \left(w\phi(x;\sigma_1^2) + (1-w)\phi(x;\sigma_2^2)\right) \log \frac{\left(w\phi(x;\sigma_1^2) + (1-w)\phi(x;\sigma_2^2)\right)}{\phi(x;1)} \mathrm{d}x$$

$$= \arg\min_{w\in[0,1]} w\frac{\sigma_1^2}{2} - w\frac{\sigma_2^2}{2} + \int \left(w\phi(x;\sigma_1^2) + (1-w)\phi(x;\sigma_2^2)\right) \log\left(w\phi(x;\sigma_1^2) + (1-w)\phi(x;\sigma_2^2)\right) \mathrm{d}x,$$

and taking first and second derivatives,

$$\frac{\mathrm{d}}{\mathrm{d}w} = \frac{\sigma_1^2 - \sigma_2^2}{2} + \int \left(\phi(x;\sigma_1^2) - \phi(x;\sigma_2^2)\right) \log\left(w\phi(x;\sigma_1^2) + (1-w)\phi(x;\sigma_2^2)\right) \mathrm{d}x$$

$$\frac{\mathrm{d}^2}{\mathrm{d}w^2} = \int \frac{\left(\phi(x;\sigma_1^2) - \phi(x;\sigma_2^2)\right)^2}{w\phi(x;\sigma_1^2) + (1-w)\phi(x;\sigma_2^2)} \mathrm{d}x > 0$$

At $w = 1$, $\int \left(\phi(x;\sigma_1^2) - \phi(x;\sigma_2^2)\right) \log\left(w\phi(x;\sigma_1^2) + (1-w)\phi(x;\sigma_2^2)\right) \mathrm{d}x = (\sigma_2^2 - \sigma_1^2)/(2\sigma_1^2)$. Therefore, if $\sigma_2^2 > \sigma_1^2 > 1$,

$$\frac{\mathrm{d}}{\mathrm{d}w} < \frac{\sigma_1^2 - \sigma_2^2}{2} + \frac{\sigma_2^2 - \sigma_1^2}{2\sigma_1^2} < 0.$$

In other words, the derivative is always negative, so the optimization sets $w = 1$ and forgets the new component. This situation occurs if $\sigma_1^2 = \tau^2 > 1$, $\sigma_2^2 = r_2\frac{\tau^2}{\tau^2-1}$ and $r_2 > \tau^2 - 1$. □

Proof of Proposition 2. Using the notation from the proof of Proposition 1, Eq. (1) is

$$\sigma^{\star 2} = \arg\min_{\sigma^2} \int \phi(x;\sigma^2) \log \frac{\phi(x;\sigma^2)^{r_1}}{\mathrm{Cauchy}(x;0,1)} \mathrm{d}x$$

$$= \arg\min_{\sigma^2} -\frac{1}{2} r_1 \log\sigma^2 + \mathbb{E}_{\mathcal{N}(0,1)}\left[\log(1+\sigma^2 x^2)\right].$$

Taking the derivative with respect to σ^2 followed by Jensen's inequality yields

$$\frac{\mathrm{d}}{\mathrm{d}\sigma^2} = \sigma^{-2}\left(-\frac{1}{2}r_1 + \mathbb{E}_{\mathcal{N}(0,1)}\left[\frac{\sigma^2 x^2}{1+\sigma^2 x^2}\right]\right)$$

$$\leq \sigma^{-2}\left(-\frac{1}{2}r_1 + \frac{\sigma^2}{1+\sigma^2}\right).$$

Therefore if $r_1 \geq 2$, the derivative with respect to σ^2 is always negative, so σ^2 increases without bound. □

B.2 Proofs of Hellinger distance properties

Proof of Proposition 4. This follows from

$$\mathrm{D}_\mathrm{H}^2(p,q) = \frac{1}{2}\int (f(x) - g(x))^2 \mu(\mathrm{d}x) \leq \frac{1}{2}\int |f(x) - g(x)|\,(f(x) + g(x))\mu(\mathrm{d}x)$$

$$= \frac{1}{2}\int \left|f^2(x) - g^2(x)\right| \mu(\mathrm{d}x) = \mathrm{D}_\mathrm{TV}(p,q)$$

and

$$\mathrm{D}_\mathrm{TV}(p,q) = \frac{1}{2}\int \left|f^2(x) - g^2(x)\right| \mu(\mathrm{d}x)$$

$$= \frac{1}{2}\int |f(x) - g(x)|(f(x) + g(x))\mu(\mathrm{d}x)$$

$$\leq \frac{1}{2}\sqrt{\int |f(x) - g(x)|^2 \mu(\mathrm{d}x) \int (f(x) + g(x))^2 \mu(\mathrm{d}x)}$$

$$= \frac{1}{\sqrt{2}}\mathrm{D}_\mathrm{H}(p,q)\sqrt{2 + 2\int f(x)g(x)\mu(\mathrm{d}x)}$$

$$= \mathrm{D}_\mathrm{H}(p,q)\sqrt{2 - \mathrm{D}_\mathrm{H}^2(p,q)}.$$

□

Proof of Proposition 5. Combining a bound on the ℓ-Wasserstein distance [51, Theorem 6.15],

$$W_\ell^\ell(p,q) \le 2^{\ell-1} \int d(x_0,x)^\ell \, |p(x) - q(x)| \, \mu(\mathrm{d}x),$$

with $|p(x) - q(x)| = |\sqrt{p(x)} - \sqrt{q(x)}|(\sqrt{p(x)} + \sqrt{q(x)})$, Cauchy-Schwarz, and Proposition 4 implies

$$W_\ell^\ell(p,q) \le 2^{\ell-1/2} \mathrm{D}_{\mathrm{H}}(p,q) \sqrt{\int d(x_0,x)^{2\ell} \left(\sqrt{p(x)} + \sqrt{q(x)}\right)^2 \mu(\mathrm{d}x)}.$$

Finally, since $(a+b)^2 \le 2a^2 + 2b^2$ for $a, b \in \mathbb{R}$,

$$W_\ell^\ell(p,q) \le 2^{\ell} \mathrm{D}_{\mathrm{H}}(p,q) \sqrt{\int d(x_0,x)^{2\ell} (p(x) + q(x)) \mu(\mathrm{d}x)}.$$

□

Proof of Proposition 6. Rearranging the definition of Hellinger distance squared,

$$\begin{aligned}
\mathrm{D}_{\mathrm{H}}^2(p,q) &= \frac{1}{2} \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mu(\mathrm{d}x) \\
&= \frac{1}{2} \int p(x) \frac{q(x)}{p(x)} \left(\sqrt{\frac{p(x)}{q(x)}} - 1\right)^2 \mu(\mathrm{d}x).
\end{aligned}$$

For $x > 1$, $x^{-1}(\sqrt{x} - 1)^2 \ge \left(\frac{\log x}{1 + \log x}\right)^2$, and for $x \le 1$, $x^{-1}(\sqrt{x} - 1)^2 \ge (\log x)^2$, so

$$\mathrm{D}_{\mathrm{H}}^2(p,q) \ge \frac{1}{2} \int_{p>q} p(x) \left(\frac{\log \frac{p(x)}{q(x)}}{1 + \log \frac{p(x)}{q(x)}}\right)^2 \mu(\mathrm{d}x) + \frac{1}{2} \int_{p \le q} p(x) \left(\log \frac{p(x)}{q(x)}\right)^2 \mu(\mathrm{d}x).$$

Now using the relation $2a^2 + 2b^2 \ge (a+b)^2$,

$$\begin{aligned}
\mathrm{D}_{\mathrm{H}}^2(p,q) &\ge \frac{1}{4} \int p(x) \left(\mathbb{1}[p > q] \frac{\log \frac{p(x)}{q(x)}}{1 + \log \frac{p(x)}{q(x)}} + \mathbb{1}[p \le q] \log \frac{p(x)}{q(x)}\right)^2 \mu(\mathrm{d}x) \\
&= \frac{1}{4} \int p(x) \left(\frac{\mathbb{1}[p > q] \log \frac{p(x)}{q(x)} + \mathbb{1}[p \le q] \log \frac{p(x)}{q(x)} \left(1 + \log \frac{p(x)}{q(x)}\right)}{1 + \log \frac{p(x)}{q(x)}}\right)^2 \mu(\mathrm{d}x) \\
&= \frac{1}{4} \int p(x) \left(\log \frac{p(x)}{q(x)}\right)^2 \left(\frac{1 + \mathbb{1}\left[\log \frac{p(x)}{q(x)} \le 0\right] \log \frac{p(x)}{q(x)}}{1 + \log \frac{p(x)}{q(x)}}\right)^2 \mu(\mathrm{d}x).
\end{aligned}$$

This provides the first result. Using the reverse Hölder inequality $\|fg\|_1 \ge \|f\|_{\frac{1}{p}} \|g\|_{\frac{-1}{p-1}}$ for $p = 2 \in (1, \infty)$,

$$\begin{aligned}
\mathrm{D}_{\mathrm{H}}^2(p,q) &\ge \frac{1}{4} \left(\int p(x) \log \frac{p(x)}{q(x)} \mu(\mathrm{d}x)\right)^2 \left(\int p(x) \left(\frac{1 + \log \frac{p(x)}{q(x)}}{1 + \mathbb{1}\left[\log \frac{p(x)}{q(x)} \le 0\right] \log \frac{p(x)}{q(x)}}\right)^2 \mu(\mathrm{d}x)\right)^{-1} \\
&= \frac{1}{4} \mathrm{D}_{\mathrm{KL}}^2(p||q) \left(\mathbb{P}\left(\log \frac{p(x)}{q(x)} \le 0\right) + \int p(x) \mathbb{1}\left[\log \frac{p(x)}{q(x)} > 0\right] \left(1 + \log \frac{p(x)}{q(x)}\right)^2 \mu(\mathrm{d}x)\right)^{-1} \\
&\ge \frac{1}{4} \mathrm{D}_{\mathrm{KL}}^2(p||q) \left(1 + \int p(x) \mathbb{1}\left[\log \frac{p(x)}{q(x)} > 0\right] \left(1 + \log \frac{p(x)}{q(x)}\right)^2 \mu(\mathrm{d}x)\right)^{-1}.
\end{aligned}$$

□

Proof of Proposition 7. This proof uses a technique adapted from [54, Theorem 1.1]. Let $Y \sim p(y)\mu(\mathrm{d}y)$, $X \sim q(x)\mu(\mathrm{d}x)$, and for $a \ge 0$,

$$\rho(x) := \left|1 - \sqrt{\frac{q(x)}{p(x)}}\right|^2 \qquad h(x) := \phi(x) \mathbb{1}[\rho(x) \le a].$$

Then by Cauchy-Schwarz,

$$\mathbb{E}\left[|I_n(\phi) - I_n(h)|\right] \leq \|\phi\|_{L^2(p)} \sqrt{\mathbb{P}\left(\rho(Y) > a\right)} \tag{11}$$

$$|I(\phi) - I(h)| \leq \|\phi\|_{L^2(p)} \sqrt{\mathbb{P}\left(\rho(Y) > a\right)} \tag{12}$$

$$\mathbb{E}\left[|I_n(h) - I(h)|\right] \leq \sqrt{N}^{-1} \sqrt{\operatorname{Var} \frac{p(X)}{q(X)} \phi(X) \mathbb{1}\left[\rho(X) \leq a\right]}. \tag{13}$$

Now note that

$$\begin{aligned}\operatorname{Var}\left(\frac{p(X)}{q(X)}\phi(X)\mathbb{1}\left[\rho(X) \leq a\right]\right) &\leq \mathbb{E}\left[\frac{p^2(X)}{q^2(X)}\phi(X)^2\mathbb{1}\left[\rho(X) \leq a\right]\right] \\ &= \mathbb{E}\left[\frac{p(Y)}{q(Y)}\phi(Y)^2\mathbb{1}\left[\rho(Y) \leq a\right]\right]\end{aligned}$$

and for $a \in [0, 1)$, $\rho(x) \leq a$ implies

$$\sqrt{\frac{q(x)}{p(x)}} \geq 1 - \sqrt{a} \implies \frac{p(x)}{q(x)} \leq (1 - \sqrt{a})^{-2}$$

So

$$\operatorname{Var}\left(\frac{p(X)}{q(X)}\phi(X)\mathbb{1}\left[\rho(X) \leq a\right]\right) \leq \|\phi\|^2_{L^2(p)} \left(\frac{1}{(1-\sqrt{a})^2}\right)$$

and hence

$$\mathbb{E}\left[|I_n(h) - I(h)|\right] \leq \|\phi\|_{L^2(p)} \sqrt{N}^{-1} \frac{1}{1 - \sqrt{a}}$$

By Markov's inequality,

$$\mathbb{P}\left(\rho(Y) > a\right) \leq a^{-1}\mathbb{E}\left[\rho(Y)\right] = 2a^{-1}\mathrm{D}_{\mathrm{H}}^2\left(p, q\right).$$

So substituting and combining the three bounds from Eqs. (11) to (13) using the triangle inequality,

$$\mathbb{E}\left[|I_n(\phi) - I(\phi)|\right] \leq \|\phi\|_{L^2(p)} \left(\frac{1}{\sqrt{N}(1-\sqrt{a})} + \sqrt{8a^{-1}}\mathrm{D}_{\mathrm{H}}\left(p, q\right)\right).$$

Optimizing over a yields

$$\sqrt{a} = \frac{8^{1/4}\mathrm{D}_{\mathrm{H}}\left(p, q\right)^{1/2}}{8^{1/4}\mathrm{D}_{\mathrm{H}}\left(p, q\right)^{1/2} + N^{-1/4}},$$

and substituting with $8^{1/4} \leq 2$ yields

$$\mathbb{E}\left[|I_n(\phi) - I(\phi)|\right] \leq \|\phi\|_{L^2(p)} \left(N^{-1/4} + 2\sqrt{\mathrm{D}_{\mathrm{H}}\left(p, q\right)}\right)^2.$$

Setting $N = \alpha^{-4}\mathrm{D}_{\mathrm{H}}\left(p, q\right)^{-2}$ yields the first result. For the second, note that $|I_n(\phi) - I(\phi)| \leq \|\phi\|_{L^2(p)}\delta$ and $|I_n(1) - 1| \leq \eta$ implies that

$$\begin{aligned}|J_n(\phi) - I(\phi)| = \frac{|I_n(\phi) - I_n(1)I(\phi)|}{I_n(1)} &\leq \frac{|I_n(\phi) - I(\phi)| + I(\phi)\,|I_n(1) - 1|}{1 - |I_n(1) - 1|} \\ &\leq \|\phi\|_{L^2(p)} \frac{\delta + \eta}{1 - \eta},\end{aligned}$$

so

$$\mathbb{P}\left(|J_n(\phi) - I(\phi)| > \|\phi\|_{L^2(p)} \frac{\delta + \eta}{1 - \eta}\right) \leq \mathbb{P}\left(|I_n(\phi) - I(\phi)| > \|\phi\|_{L^2(p)}\delta\right) + \mathbb{P}\left(|I_n(\phi) - 1| > \eta\right),$$

which by Markov inequality and the previous bound,

$$\mathbb{P}\left(|J_n(\phi) - I(\phi)| > \|\phi\|_{L^2(p)} \frac{\delta + \eta}{1 - \eta}\right) \leq \left(N^{-1/4} + 2\sqrt{\mathrm{D}_{\mathrm{H}}\left(p, q\right)}\right)^2 \left(\delta^{-1} + \eta^{-1}\right)$$

Minimizing $\delta^{-1} + \eta^{-1}$ subject to the constraint that $t = (\delta + \eta)/(1 - \eta)$ yields the result. □

Proof of Proposition 8. For the first bound, by Jensen's inequality

$$\begin{aligned}\mathbb{E}\left[\left|\widehat{\mathrm{D}_{\mathrm{H}}^2(p,q)} - \mathrm{D}_{\mathrm{H}}^2(p,q)\right|\right] &\leq \sqrt{\mathrm{Var}\,\frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{p(X_n)}{q(X_n)}}} \\ &= \sqrt{\frac{1}{N}\,\mathrm{Var}\sqrt{\frac{p(X_n)}{q(X_n)}}} \\ &= \sqrt{\frac{1}{N}}\sqrt{1-\left(\int\sqrt{q(x)p(x)}\,\mathrm{d}x\right)^2} \\ &= \sqrt{\frac{1}{N}}\sqrt{\mathrm{D}_{\mathrm{H}}^2(p,q)\left(2-\mathrm{D}_{\mathrm{H}}^2(p,q)\right)}.\end{aligned}$$

For the second bound, using the triangle inequality, and cancelling out normalization constants

$$\begin{aligned}&\mathbb{E}\left[\left|\widetilde{\mathrm{D}_{\mathrm{H}}^2(p,q)} - \mathrm{D}_{\mathrm{H}}^2(p,q)\right|\right] \\ \leq&\mathbb{E}\left[\frac{\frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{p(X_n)}{q(X_n)}}}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}\frac{p(X_n)}{q(X_n)}}}\left|1-\sqrt{\frac{\frac{1}{N}\sum_{n=1}^{N}\frac{p(X_n)}{q(X_n)}}{\mathbb{E}\left[\frac{p(X_n)}{q(X_n)}\right]}}\right|\right] + \mathbb{E}\left[\left|\frac{\frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{p(X_n)}{q(X_n)}} - \mathbb{E}\left[\sqrt{\frac{p(X_n)}{q(X_n)}}\right]}{\sqrt{\mathbb{E}\left[\frac{p(X_n)}{q(X_n)}\right]}}\right|\right]\end{aligned}$$

By Jensen's inequality on the left term and Cauchy-Schwarz on the right, and noting that $\mathbb{E}[p/q] = 1$,

$$\mathbb{E}\left[\left|\widetilde{\mathrm{D}_{\mathrm{H}}^2(p,q)} - \mathrm{D}_{\mathrm{H}}^2(p,q)\right|\right] \leq \mathbb{E}\left[\left|1-\sqrt{\frac{1}{N}\sum_{n=1}^{N}\frac{p(X_n)}{q(X_n)}}\right|\right] + \sqrt{\frac{2}{N}}\,\mathrm{D}_{\mathrm{H}}(p,q)$$

The left term can be bounded via Cauchy-Schwarz and Jensen's inequality:

$$\begin{aligned}\mathbb{E}\left[\left|1-\sqrt{\frac{1}{N}\sum_{n=1}^{N}\frac{p(X_n)}{q(X_n)}}\right|\right] &\leq \sqrt{2-2\mathbb{E}\left[\sqrt{\frac{1}{N}\sum_{n=1}^{N}\frac{p(X_n)}{q(X_n)}}\right]} \\ &\leq \sqrt{2-2\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{p(X_n)}{q(X_n)}}\right]} \\ &= \sqrt{2}\,\mathrm{D}_{\mathrm{H}}(p,q)\end{aligned}$$

Combining these results yields the second inequality. □

B.3 Theoretical tools for establishing convergence of Algorithm 1

Lemma 9. *Define* $\hat{f} := \arg\min_{h\in \mathrm{cl\,span}\,\mathcal{H}:\|h\|_2=1}\|f-h\|_2$. *Then* $\hat{f}$ *exists, is unique, and is nonnegative.*

Proof of Lemma 9. Since $\mathrm{cl\,span}\,\mathcal{H}$ is a closed convex set, there exists a unique function $\hat{f}'$ of minimum distance to f. Note that $\hat{f}'$ is nonnegative since f is nonnegative, so otherwise $\hat{f}'$ could be replaced with $\max\{0,\hat{f}'\}$ without increasing the distance to f. Furthermore, the error $\epsilon := f - \hat{f}'$ is orthogonal to $\mathrm{cl\,span}\,\mathcal{H}$. Since f is not orthogonal to $\mathrm{cl\,span}\,\mathcal{H}$, $\hat{f}' \neq 0$, so set $\hat{f} = \frac{\hat{f}'}{\|\hat{f}'\|_2}$. Suppose there is another unit-norm function $g \in \mathrm{cl\,span}\,\mathcal{H}$ at least as close to f; then

$$\begin{aligned}0 \geq \left\langle f, \frac{\hat{f}'}{\|\hat{f}'\|_2} - g\right\rangle = \left\langle \hat{f}' + \epsilon, \frac{\hat{f}'}{\|\hat{f}'\|_2} - g\right\rangle &= \left\langle \hat{f}', \frac{\hat{f}'}{\|\hat{f}'\|_2} - g\right\rangle \\ &= \|\hat{f}'\|_2 - \left\langle \hat{f}', g\right\rangle.\end{aligned}$$

Dividing both sides by $\|\hat{f}'\|_2$ yields the inequality $\left\langle \hat{f}, g\right\rangle \geq 1$, implying that $g = \hat{f}$, and thus $\hat{f}$ is unique. □

Lemma 10. $\tau \leq \frac{\left\langle \hat{f}, g_1\right\rangle}{1-\sqrt{1-\left\langle \hat{f}, g_1\right\rangle^2}} < \infty$.

Proof of Lemma 10. Set $h_1 = \left\langle \hat{f}, g_1 \right\rangle g_1$ where g_1 is chosen from Eq. (5), and $\forall\, i > 1$, set $h_i = 0$. Since f is not orthogonal to $\operatorname{cl}\operatorname{span}\mathcal{H}$, $\left\langle \hat{f}, g_1 \right\rangle > 0$, so $\tau < \infty$. □

Lemma 11. *Suppose at each iteration, the optimization in Eq.* (5) *is solved with multiplicative error* $(1-\delta)$ *relative to the optimal objective. Then*

$$J_{n+1} \leq J_n(1 - J_n) \text{ where } J_n := \left(\frac{1-\delta}{\tau}\right)^2 \left(1 - \left\langle \hat{f}, \bar{g}_n \right\rangle^2\right).$$

Proof of Lemma 11. Taking the derivative of the objective in Eq. (4) with respect to x and setting to zero, the solution is

$$x^\star = \sqrt{\frac{\left\langle f, \frac{h - \langle h, \bar{g}_n\rangle \bar{g}_n}{\|h - \langle h, \bar{g}_n\rangle \bar{g}_n\|_2} \right\rangle^2}{\left\langle f, \frac{h - \langle h, \bar{g}_n\rangle \bar{g}_n}{\|h - \langle h, \bar{g}_n\rangle \bar{g}_n\|_2} \right\rangle^2 + \langle f, \bar{g}_n\rangle^2}}. \tag{14}$$

Suppose at iteration $n+1$, instead of g_{n+1} we obtain a function h satisfying a $(1-\delta)$-relative approximation to Eq. (5). Then using the optimal value for $x^\star$ from Eq. (14), noting that the quadratic weight optimization provides at least as much error reduction as the geodesic update with $x^\star$, and noting that $f = \hat{f}' + \epsilon$ where $\epsilon \perp \operatorname{cl}\operatorname{span}\mathcal{H}$ and $\hat{f}'$ is from the proof of Lemma 9, we find the recursion

$$\begin{aligned}
&\|\hat{f}'\|_2^2 - \left\langle \hat{f}', \bar{g}_{n+1} \right\rangle^2 \\
&= \left(\|\hat{f}'\|_2^2 - \left\langle \hat{f}', \bar{g}_n \right\rangle^2\right) \left(1 - \left\langle \frac{\hat{f}' - \left\langle \hat{f}', \bar{g}_n \right\rangle \bar{g}_n}{\left\| \hat{f}' - \left\langle \hat{f}', \bar{g}_n \right\rangle \bar{g}_n \right\|}, \frac{h - \langle h, \bar{g}_n\rangle \bar{g}_n}{\|h - \langle h, \bar{g}_n\rangle \bar{g}_n\|} \right\rangle^2\right) \\
&\leq \left(\|\hat{f}'\|_2^2 - \left\langle \hat{f}', \bar{g}_n \right\rangle^2\right) \left(1 - (1-\delta)^2 \left\langle \frac{\hat{f}' - \left\langle \hat{f}', \bar{g}_n \right\rangle \bar{g}_n}{\left\| \hat{f}' - \left\langle \hat{f}', \bar{g}_n \right\rangle \bar{g}_n \right\|}, \frac{g_{n+1} - \langle g_{n+1}, \bar{g}_n\rangle \bar{g}_n}{\|g_{n+1} - \langle g_{n+1}, \bar{g}_n\rangle \bar{g}_n\|} \right\rangle^2\right).
\end{aligned}$$

Now again using the fact that $\epsilon \perp \operatorname{cl}\operatorname{span}\mathcal{H}$ as well as the fact that g_{n+1} is the argmax of Eq. (5), we can replace g_{n+1} with any convex combination of other elements of $\mathcal{H}$, so

$$\begin{aligned}
&\|\hat{f}'\|_2^2 - \left\langle \hat{f}', \bar{g}_{n+1} \right\rangle^2 \\
&\leq \left(\|\hat{f}'\|_2^2 - \left\langle \hat{f}', \bar{g}_n \right\rangle^2\right) \left(1 - (1-\delta)^2 \left\langle \frac{f - \langle f, \bar{g}_n\rangle \bar{g}_n}{\left\| \hat{f}' - \left\langle \hat{f}', \bar{g}_n \right\rangle \bar{g}_n \right\|}, \frac{g_{n+1} - \langle g_{n+1}, \bar{g}_n\rangle \bar{g}_n}{\|g_{n+1} - \langle g_{n+1}, \bar{g}_n\rangle \bar{g}_n\|} \right\rangle^2\right) \\
&\leq \left(\|\hat{f}'\|_2^2 - \left\langle \hat{f}', \bar{g}_n \right\rangle^2\right) \left(1 - (1-\delta)^2 \sup_{h_i \in \operatorname{cone}\mathcal{H}} \left\langle \frac{f - \langle f, \bar{g}_n\rangle \bar{g}_n}{\left\| \hat{f}' - \left\langle \hat{f}', \bar{g}_n \right\rangle \bar{g}_n \right\|}, \frac{\sum_i h_i - \left\langle \sum_i h_i, \bar{g}_n\right\rangle \bar{g}_n}{D} \right\rangle^2\right),
\end{aligned}$$

where $D = \sum_i \|h_i\| \, \|h_i - \langle h_i, \bar{g}_n\rangle \, \bar{g}_n\|$. Define $\nu := \sum_i h_i - \hat{f}$. Again using $\epsilon \perp \operatorname{cl}\operatorname{span}\mathcal{H}$, and normalizing the left vector by $\|\hat{f}'\|$ yields

$$= \left(\|\hat{f}'\|_2^2 - \left\langle \hat{f}', \bar{g}_n \right\rangle^2\right) \left(1 - (1-\delta)^2 \sup_{h_i \in \operatorname{cone}\mathcal{H}} \left\langle \frac{\hat{f} - \left\langle \hat{f}, \bar{g}_n \right\rangle \bar{g}_n}{\left\| \hat{f} - \left\langle \hat{f}, \bar{g}_n \right\rangle \bar{g}_n \right\|}, \frac{\hat{f} - \left\langle \hat{f}, \bar{g}_n \right\rangle \bar{g}_n + \nu - \langle \nu, \bar{g}_n\rangle \bar{g}_n}{D} \right\rangle^2\right).$$

Now noting that the inner term is minimized when $\nu = -\|\nu\| \hat{f}$, we have that

$$\leq \left(\|\hat{f}'\|_2^2 - \left\langle \hat{f}', \bar{g}_n \right\rangle^2\right) \left(1 - \sup_{h_i \in \operatorname{cone}\mathcal{H}} \frac{(1-\delta)^2 (1 - \|\nu\|)^2}{D^2} (1 - \left\langle \hat{f}, \bar{g}_n \right\rangle^2)\right).$$

Finally, dividing both sides by $\|\hat{f}'\|_2^2$ and noting that $D \leq \sum_i \|h_i\|$,

$$1 - \left\langle \hat{f}, \bar{g}_{n+1} \right\rangle^2 \leq \left(1 - \left\langle \hat{f}, \bar{g}_n \right\rangle^2\right) \left(1 - \left(\frac{1-\delta}{\tau}\right)^2 \left(1 - \left\langle \hat{f}, \bar{g}_n \right\rangle^2\right)\right).$$

Denoting $J_n = \left(\frac{1-\delta}{\tau}\right)^2 \left(1 - \left\langle \hat{f}, \bar{g}_n \right\rangle^2\right)$, and multiplying both sides by $\left(\frac{1-\delta}{\tau}\right)^2$ yields the recursion

$$J_{n+1} \leq J_n \left(1 - J_n\right).$$

□

Proof of Theorem 3. By [59, Lemma A.6], the recursion Lemma 11 satisfies $J_n \leq \frac{J_0}{1+J_0 n}$. Substituting the definition of J_n and noting that $\mathrm{D_H}\left(\hat{p}, q_n\right)^2 = 1 - \left\langle \hat{f}, \bar{g}_n \right\rangle \leq 1 - \left\langle \hat{f}, \bar{g}_n \right\rangle^2$ yields the result. □