point $f_0 \in \mathcal{F}$ can be viewed as an element of $\mathcal{F}^*$, which we write $\partial_f^{in} C|_{f_0}$. We denote by $d|_{f_0} \in \mathcal{F}$, a corresponding dual element, such that $\partial_f^{in} C|_{f_0} = \langle d|_{f_0}, \cdot \rangle_{p^{in}}$.

The *kernel gradient* $\nabla_K C|_{f_0} \in \mathcal{F}$ is defined as $\Phi_K \left(\partial_f^{in} C|_{f_0} \right)$. In contrast to $\partial_f^{in} C$ which is only defined on the dataset, the kernel gradient generalizes to values x outside the dataset thanks to the kernel K:

$$\nabla_K C|_{f_0}(x) = \frac{1}{N} \sum_{j=1}^{N} K(x, x_j) d|_{f_0}(x_j).$$

A time-dependent function $f(t)$ follows the *kernel gradient descent with respect to* K if it satisfies the differential equation

$$\partial_t f(t) = -\nabla_K C|_{f(t)}.$$

During kernel gradient descent, the cost $C(f(t))$ evolves as

$$\partial_t C|_{f(t)} = - \left\langle d|_{f(t)}, \nabla_K C|_{f(t)} \right\rangle_{p^{in}} = - \left\| d|_{f(t)} \right\|_K^2 .$$

Convergence to a critical point of C is hence guaranteed if the kernel K is positive definite with respect to $|| \cdot ||_{p^{in}}$: the cost is then strictly decreasing except at points such that $||d|_{f(t)}||_{p^{in}} = 0$. If the cost is convex and bounded from below, the function $f(t)$ therefore converges to a global minimum as $t \to \infty$.

3.1 Random functions approximation

As a starting point to understand the convergence of ANN gradient descent to kernel gradient descent in the infinite-width limit, we introduce a simple model, inspired by the approach of (19).

A kernel K can be approximated by a choice of P random functions $f^{(p)}$ sampled independently from any distribution on $\mathcal{F}$ whose (non-centered) covariance is given by the kernel K:

$$\mathbb{E}[f_k^{(p)}(x) f_{k'}^{(p)}(x')] = K_{kk'}(x, x').$$

These functions define a random linear parametrization $F^{lin} : \mathbb{R}^P \to \mathcal{F}$

$$\theta \mapsto f_\theta^{lin} = \frac{1}{\sqrt{P}} \sum_{p=1}^{P} \theta_p f^{(p)}.$$

The partial derivatives of the parametrization are given by

$$\partial_{\theta_p} F^{lin}(\theta) = \frac{1}{\sqrt{P}} f^{(p)}.$$

Optimizing the cost $C \circ F^{lin}$ through gradient descent, the parameters follow the ODE:

$$\partial_t \theta_p(t) = -\partial_{\theta_p} (C \circ F^{lin})(\theta(t)) = -\frac{1}{\sqrt{P}} \partial_f^{in} C|_{f_{\theta(t)}^{lin}} \; f^{(p)} = -\frac{1}{\sqrt{P}} \left\langle d|_{f_{\theta(t)}^{lin}}, f^{(p)} \right\rangle_{p^{in}} .$$

As a result the function $f_{\theta(t)}^{lin}$ evolves according to

$$\partial_t f_{\theta(t)}^{lin} = \frac{1}{\sqrt{P}} \sum_{p=1}^{P} \partial_t \theta_p(t) f^{(p)} = -\frac{1}{P} \sum_{p=1}^{P} \left\langle d|_{f_{\theta(t)}^{lin}}, f^{(p)} \right\rangle_{p^{in}} f^{(p)},$$

where the right-hand side is equal to the kernel gradient $-\nabla_{\tilde{K}} C$ with respect to the *tangent kernel*

$$\tilde{K} = \sum_{p=1}^{P} \partial_{\theta_p} F^{lin}(\theta) \otimes \partial_{\theta_p} F^{lin}(\theta) = \frac{1}{P} \sum_{p=1}^{P} f^{(p)} \otimes f^{(p)}.$$

This is a random n_L-dimensional kernel with values $\tilde{K}_{ii'}(x, x') = \frac{1}{P} \sum_{p=1}^{P} f_i^{(p)}(x) f_{i'}^{(p)}(x')$.

Performing gradient descent on the cost $C \circ F^{lin}$ is therefore equivalent to performing kernel gradient descent with the tangent kernel $\tilde{K}$ in the function space. In the limit as $P \to \infty$, by the law of large numbers, the (random) tangent kernel $\tilde{K}$ tends to the fixed kernel K, which makes this method an approximation of kernel gradient descent with respect to the limiting kernel K.