Topics in Applied Mathematics: Infinitely Large Neural Networks, 3341.751
E. Ryu
Spring 2022

Homework 5
Due 5pm, Friday, May 20, 2022

Problem 1: *Preconditioned gradient flow.* Let $\mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}$ be a differentiable convex function. Assume a minimizer $\theta_\star$ exists. Let $M \in \mathbb{R}^{p\times p}$ be symmetric (strictly) positive definite. Consider the *preconditioned gradient flow*

$$\dot{\theta}(t) = -M\nabla\mathcal{L}(\theta(t)), \qquad \theta(0) = \theta_0.$$

Show that (i) $\frac{d}{dt}\mathcal{L}(\theta(t)) \le 0$ for all $t > 0$ and (ii) $\mathcal{L}(\theta(t)) \to \mathcal{L}(\theta_\star)$ as $t \to \infty$.

Remark. Applying a positive definite matrix to the gradient is referred to as "preconditioning", since the right choice of M can reduce the "condition number" and accelerate convergence. In fact, $M = (\nabla^2\mathcal{L}(\theta))^{-1}$ corresponds to Newton's method.

Problem 2: *Variational formulation of gradient flow.* Assume that $\mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}$ is differentiable and that $\nabla\mathcal{L}\colon \mathbb{R}^p \to \mathbb{R}^p$ is L-Lipschitz continuous and M-bounded. For $\alpha > 0$, define the sequence $\{\theta^k_{(\alpha)}\}_{k\in\mathbb{N}}$ as

$$\theta^{k+1}_{(\alpha)} = \operatorname*{argmin}_{\theta\in\mathbb{R}^p} \left\{ \mathcal{L}(\theta) + \frac{1}{2\alpha}\|\theta - \theta^k_{(\alpha)}\|^2 \right\},$$

with $\theta^0_{(\alpha)} = \theta^0 \in \mathbb{R}^p$. Assume that the argmin uniquely exists. Let $\theta(t)$ be the gradient flow starting from $\theta(0) = \theta^0$. Show that for any $T < \infty$,

$$\sup_{t\in[0,T]} \|\theta(t) - \theta^{\lfloor t/\alpha \rfloor}_{(\alpha)}\| \to 0$$

as $\alpha \to 0$.

Remark. We say $\mathcal{L}\colon \mathbb{R}^d \to \mathbb{R}$ is λ-semiconvex if $\mathcal{L}(\theta) + (\lambda/2)\|\theta\|^2$ is a convex function. If $\mathcal{L}$ is λ-semiconvex, then $\{\theta^k_{(\alpha)}\}_{k\in\mathbb{N}}$ is well defined for $\alpha < 1/\lambda$.

Problem 3: *Matrix-valued PDK from vector-valued features.* Let $\phi\colon \mathcal{X} \to \mathbb{R}^{d\times M}$ and write

$$\phi(x) = \begin{bmatrix}\psi_1(x) & \psi_2(x) & \cdots & \psi_M(x)\end{bmatrix}.$$

Assume $\psi_1, \dots, \psi_M \colon \mathcal{X} \to \mathbb{R}^d$ are linearly independent as functions. Consider the mvPDK $K\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}^{d\times d}$ defined as

$$K(x, x') = (\phi(x))(\phi(x'))^\intercal$$

or equivalently,

$$K = \sum_{k=1}^{M} \psi_k \otimes \psi_k.$$

Let

$$\mathcal{H} = \operatorname{span}\{\psi_k\}_{k=1}^{M}.$$

For

$$f = \sum_{k=1}^{M} \alpha_k \psi_k, \qquad g = \sum_{k=1}^{M} \beta_k \psi_k,$$

define the inner product

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{k=1}^{M} \alpha_k \beta_k.$$

Show that $\mathcal{H}$ is the vvRKHS corresponding to K.

Problem 4: *Eigenfunctions of L_K with respect to a finitely-supported measure.* Let $\mathcal{X}$ be a nonempty set. Let $\mathbb{R}^{\mathcal{X}}$ denote the set of functions $f\colon \mathcal{X} \to \mathbb{R}$. Let $K\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a strictly positive definite kernel. Let $X_1, \dots, X_N \in \mathcal{X}$ be distinct. Consider the operator $L_K\colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ defined as

$$L_K[f] = \sum_{i=1}^{N} K(\cdot, X_i) f(X_i).$$

Let $G \in \mathbb{R}^{N\times N}$ be the kernel matrix defined as $G_{ij} = K(X_i, X_j)$ for $i, j \in \{1, \dots, N\}$. Let $u_1, \dots, u_n$ be the orthonormal eigenvectors of G with respective positive eigenvalues $\lambda_1, \dots, \lambda_N$.

(i) Define

$$f^{(i)} = \sum_{j=1}^{N} K(\cdot, X_j)(u_i)_j, \qquad \text{for } i = 1, \dots, N.$$

Show that $f^{(i)}$ is an eigenfunction of L_K with eigenvalue λ_i, i.e., $L_K[f^{(i)}] = \lambda_i f^{(i)}$, for $i = 1, \dots, N$.

(ii) Show that

$$\mathbb{R}^{\mathcal{X}} = \underbrace{\{f \in \mathbb{R}^{\mathcal{X}} \mid f(x_i) = 0 \text{ for } i = 1, \dots, N\}}_{:=V_0} \oplus \operatorname{span}\{f^{(1)}, \dots, f^{(N)}\},$$

i.e., for any $f \in \mathbb{R}^{\mathcal{X}}$, we can find a unique decomposition

$$f = f^{(0)} + \sum_{i=1}^{N} \alpha_i f^{(i)}, \qquad f^{(0)} \in V_0.$$

(iii) Show that any $f^{(0)} \in V_0$ is an eigenfunction of L_K with eigenvalue 0.

(iv) Define $P\colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^N$ as $(P[f])_i = f(X_i)$ for $i = 1, \ldots, N$. Show that if

$$f = f^{(0)} + \sum_{i=1}^{N} \alpha_i f^{(i)}, \qquad f^{(0)} \in V_0,$$

then

$$u_i^\intercal G^{-1} P[f] = \alpha_i, \qquad \text{for } i = 1, \ldots, N.$$

(v) Consider the ordinary differential equation

$$\dot{f}_t = -L_K[f_t]$$

with initial condition f_0 at $t = 0$. Let

$$f_0 = f_0^{(0)} + \sum_{i=1}^{N} \alpha_i f^{(i)}, \qquad f_0^{(0)} \in V_0$$

be the eigenfunction expansion of f_0. Show that

$$f_t = f_0^{(0)} + \sum_{i=1}^{N} \alpha_i e^{-t\lambda} f_0^{(i)}$$

solves the differential equation.

(vi) Show that

$$\lim_{t\to\infty} f_t(x) = f_0^{(0)}(x) = f_0(x) - \sum_{j=1}^{N} K(x, X_j)(G^{-1}P[f_0])_j, \qquad \forall\, x \in \mathcal{X}.$$

Hint. For (i), use P as defined in (iv) and $P^\dagger\colon \mathbb{R}^N \to \mathbb{R}^{\mathcal{X}}$ defined as

$$(P^\dagger(v))(x) = \begin{cases} v_i & \text{if } x = X_i \text{ for any } i = 1, \ldots, N \\ 0 & \text{otherwise.} \end{cases}$$

Then $L_K = L_K P^\dagger P$, i.e.,

$$L_K[f] = L_K[P^\dagger(P[f])]$$

for all $f \in \mathbb{R}^{\mathcal{X}}$.

Remark. The solution to the ODE of (v) is often expressed via the "exponential map"

$$f_t = e^{-tL_K} f_0.$$

Problem 5: Let $\mathcal{X} \subseteq \mathbb{R}^d$ be nonempty and let $P \in \mathcal{P}(\mathcal{X})$ be a probability measure. Let $R\colon L^2(P;\mathbb{R}^k) \to \mathbb{R}$ be Fréchet differentiable everywhere with derivative $\partial R|_{f_0}\colon \mathcal{X} \to \mathbb{R}^k$ for all $f_0 \in L^2(P;\mathbb{R}^k)$. For notational simplicity, we often suppress the dependence on f_0 and write $\partial R = \partial R|_{f_0}$. Define $(\partial R)_i\colon \mathcal{X} \to \mathbb{R}$ to be the ith coordinate of ∂R, i.e., $(\partial R)_i \in L^2(P;\mathbb{R})$ and $(\partial R)_i(x) = e_i^\intercal \partial R(x)$, where $e_i \in \mathbb{R}^k$ is the ith unit vector, for $i = 1, \dots, k$.

(i) Show that
$$R[f_0 + \delta \otimes e_i] = R[f_0] + \langle (\partial R)_i|_{f_0}, \delta\rangle_{L^2(P;\mathbb{R})} + o(\|\delta\|_{L^2(P;\mathbb{R})})$$
for small $\delta \in L^2(P;\mathbb{R})$.

(ii) Assume R has the decomposition
$$R[f] = \sum_{i=1}^k R_i[f_i]$$
for any $f = (f_1, \dots, f_k) \in L^2(P;\mathbb{R}^k)$. So $f_1, \dots, f_k \in L^2(P;\mathbb{R})$ and $R_i\colon L^2(P;\mathbb{R}) \to \mathbb{R}$ for $i = 1, \dots, k$. Show that R_i is Fréchet differentiable everywhere with derivative
$$\partial(R_i) = (\partial R)_i, \qquad \text{for } i = 1, \dots, k.$$

Clarification. For all $x \in \mathcal{X}$,
$$(f_0 + \delta \otimes e_i)(x) = \begin{bmatrix} (f_0(x))_1 \\ (f_0(x))_2 \\ \vdots \\ (f_0(x))_{i-1} \\ (f_0(x))_i + \delta(x) \\ (f_0(x))_{i+1} \\ \vdots \\ (f_0(x))_k \end{bmatrix}.$$
Therefore, $(\partial_f R|_{f_0})_i$ is the derivative of R with respect the infinitestimal changes in the ith output of the input function f_0.

Problem 6: Let $\mathcal{X} \subseteq \mathbb{R}^d$ be nonempty, $X_1, \dots, X_N \in \mathcal{X}$, and
$$P = \frac{1}{N}\sum_{i=1}^N \delta_{X_i}.$$
Let $R\colon L^2(P;\mathbb{R}^k) \to \mathbb{R}$ be Fréchet differentiable everywhere with derivative $\partial R|_{f_0}\colon \mathcal{X} \to \mathbb{R}^k$ for all $f_0 \in L^2(P;\mathbb{R}^k)$. Assume $f_\theta(x)$ is differentiable in θ for all x. Show that
$$\frac{\partial}{\partial \theta_p} R[f_\theta] = \left\langle \frac{\partial f_\theta}{\partial \theta_p}, \partial_f R \right\rangle_{L^2(P;\mathbb{R}^k)}$$
or, to be more precise, that
$$\left(\frac{\partial}{\partial \theta_p} R[f_\theta]\right)\bigg|_{\theta=\theta_0} = \left\langle \frac{\partial f_\theta}{\partial \theta_p}\bigg|_{\theta=\theta_0}, \partial_f R|_{f_{\theta_0}} \right\rangle_{L^2(P;\mathbb{R}^k)}.$$

Hint. Differentiability of $f_\theta(x)$ in θ implies directional differentiability
$$f_{\theta_0 + he_i}(x) = f_{\theta_0}(x) + \frac{df_\theta(x)}{d\theta_i}h + o(h).$$

Problem 7: *General NTK calculation for MLPs.* Consider the depth-L MLP

$$
\begin{aligned}
f_\theta(x) &= y_L \\
y_L &= z_L, & z_L &= \frac{\sigma_A}{\sqrt{n_{L-1}}} A_L y_{L-1} + \sigma_b b_L \in \mathbb{R}^{n_L}, \\
y_{L-1} &= \sigma(z_{L-1}), & z_{L-1} &= \frac{\sigma_A}{\sqrt{n_{L-2}}} A_{L-1} y_{L-2} + \sigma_b b_{L-1} \in \mathbb{R}^{n_{L-1}}, \\
&\vdots \\
y_2 &= \sigma(z_2), & z_2 &= \frac{\sigma_A}{\sqrt{n_1}} A_2 y_1 + \sigma_b b_2 \in \mathbb{R}^{n_2}, \\
y_1 &= \sigma(z_1), & z_1 &= \frac{\sigma_A}{\sqrt{n_0}} A_1 x + \sigma_b b_1 \in \mathbb{R}^{n_1},
\end{aligned}
$$

where $\sigma_A > 0$, $\sigma_b > 0$, $x \in \mathbb{R}^{n_0}$, $A_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$, and $b_\ell \in \mathbb{R}^{n_\ell}$. For $\ell = 1, \dots, L$, define

$$\theta^{(\ell)} = (A_1, b_1, A_2, b_2, \dots, A_\ell, b_\ell)$$

and

$$\Theta^{(\ell)}(x, x') = \left(\frac{\partial z_\ell(x)}{\partial \theta^{(\ell)}}\right) \left(\frac{\partial z_\ell(x')}{\partial \theta^{(\ell)}}\right)^\intercal.$$

Show that

$$\Theta^{(1)}(x, x') = \left(\frac{\sigma_A^2}{n_0} x^\intercal x' + \sigma_b^2\right) I_{n_1}$$

for all $x, x' \in \mathbb{R}^{n_0}$, and that

$$
\begin{aligned}
\Theta^{(\ell+1)}(x, x') &= \left(\frac{\sigma_A^2}{n_\ell} \sigma(z_\ell(x))^\intercal \sigma(z_\ell(x')) + \sigma_b^2\right) I_{n_\ell} \\
&\quad + \frac{\sigma_A^2}{n_\ell} A_{\ell+1} \operatorname{diag}(\sigma'(z_\ell(x))) \Theta^{(\ell)}(x, x') \operatorname{diag}(\sigma'(z_\ell(x'))) A_{\ell+1}^\intercal
\end{aligned}
$$

for all $x, x' \in \mathbb{R}^{n_0}$ and $\ell = 1, \dots, L-1$.

Clarification. We do not assume $n_L = 1$. We do not take any infinite-width limits in this problem. We are not considering gradient flow or any process for updating the parameters in this problem.