TD Data Science

The following exercises are meant to illustrate Gabriel Peyré's ***Mathematics of Data*** *course (`mathematical-tours.github.io`). Some of them are adapted from the past years' exams.*

Please refer to Geert-Jan Huizing (`huizing@ens.fr`) for questions these exercises and their solutions.

Exercises

[⋆⋆] **Coordinate descent (same as last week)**

We let $f : \mathbb{R}^p \to \mathbb{R}$. Coordinate descent is an optimization method that tries to minimize f alternatively with respect to individual coordinates.

We denote w^t the iterates. At iteration t, we chose an index $i \in \{1, \dots, p\}$ and try to minimize f with respect to w_i^t without changing the other coordinates w_j^t, $j \neq i$. More formally, we define $\varphi_i(x, w) = f(w_1, \dots, w_{i-1}, x, w_{i+1}, w_p)$ and set at each iteration:

$$w_i^{t+1} = \arg\min_x \varphi_i(x, w^t) \text{ and } w_j^{t+1} = w_j^t \text{ for } j \neq i$$

The index i is typically chosen as cyclic : $i = 1 + (t \mod p)$. Therefore , at iteration 1, the coordinate 1 is updated, at iteration 2, the coordinate 2 is updated, ... , at iteration p the coordinate p is updated and at iteration $p + 1$ the coordinate 1 is modified again.

1. Assume that f is the quadratic function:

   $$f(w) = \frac{1}{2}\langle w, Aw\rangle - \langle b, w\rangle$$

   Compute the update rule to minimize φ_i.

2. At iteration $t + 1$, we update the coordinate i. Demonstrate that

   $$f(w^{t+1}) - f(w^t) = -\frac{(Aw^t - b)_i^2}{2A_{ii}} \leq -\frac{(Aw^t - b)_i^2}{2A_{\max}}$$

   where $A_{max} = \max_i A_{ii}$

3. At iteration t, the coordinate that is updated is i such that $(Aw^t - b)_i^2$ is maximal. Show that

   $$f(w^{t+1}) - f(w^t) \leq -\frac{\|Aw^t - b\|^2}{2pA_{\max}}$$

4. Let $w^* = A^{-1}b$. Demonstrate that $\|Aw - b\|^2 \geq 2\sigma_{\min}(A)(f(w) - f(w^*))$.

   Provide a convergence rate for the coordinate descent method. What is the difference with gradient descent ? When is it faster, or slower? Hint: what is the link between A_{max} and $\sigma_{\max}(A)$?

[$\star\star$] **Inverse problems**

L2 regularization We consider $f_0 \in \mathbb{R}^n$ a discrete signal. We only observe the signal $y \overset{\text{def.}}{=} f_0 \star h + \varepsilon$, where $h \in \mathbb{R}^n$ is a low-pass filter and $\varepsilon \in \mathbb{R}^n$ is noise. For some $\lambda \in \mathbb{R}_+^*$, we look for

$$\arg\min_{f \in \mathbb{R}^n} \frac{1}{2} \|f \star h - y\|^2 + \frac{\lambda}{2} \|f\|^2 .$$

1. Using the Fourier decomposition, prove that for an optimal $f \in \mathbb{R}^n$,

$$\hat{f}_k = \frac{\hat{h}_k^*}{|\hat{h}_k|^2 + \lambda} \hat{y}_k.$$

2. Why does $\lambda > 0$ improve the deconvolution in presence of noise?

Sobolev regularization For $f \in \mathbb{R}^n$, we denote $Gf = (f_i - f_{i-1})_i$ (we consider indexes modulo n).

1. What is the adjoint operator $G^\top : \mathbb{R}^n \to \mathbb{R}^n$ for the canonical inner product? *i.e.* $\langle Gf, u \rangle = \langle f, G^\top u \rangle$.
2. Show that G, $G^\top$ and $L \overset{\text{def.}}{=} GG^\top$ are discrete convolution operators and give their associated filters g, $\tilde{g}$ and ℓ.
3. Compute the discrete Fourier coefficients $\hat{g}$. Express $\hat{\tilde{g}}$ and $\hat{l}$ as a function of $\hat{g}$.
4. We now consider the problem

$$\arg\min_{f \in \mathbb{R}^n} \frac{1}{2} \|f \star h - y\|^2 + \frac{\lambda}{2} \|Gf\|^2 .$$

Using the Fourier decomposition and the results from part 2, prove that for an optimal $f \in \mathbb{R}^n$,

$$\hat{f}_k = \frac{\hat{h}_k^*}{|\hat{h}_k|^2 + \lambda \hat{\ell}_k} \hat{y}_k.$$

5. How does this expression differ from part 1?

Total Variation regularization We define $\forall x \in \mathbb{R}$, $\|x\|_\varepsilon \overset{\text{def.}}{=} \sqrt{x^2 + \varepsilon^2}$ and $\forall f \in \mathbb{R}^d$, $\|f\|_{1,\varepsilon} \overset{\text{def.}}{=} \sum_i \|f_i\|_\varepsilon$.

For a linear operator H, we recall the chain rule: $\nabla(\varphi \circ H)(f) = H^\top (\nabla \varphi)(Hf)$ where $H^\top$ is the adjoint of H.

1. Compute $\nabla \|\cdot\|_{1,\varepsilon}$.
2. Compute the gradient of $J_{\text{TV}}^\varepsilon : f \mapsto \|Gf\|_{1,\varepsilon}$.
3. We now consider the deconvolution problem with ε-smoothed TV regularization:

$$\arg\min_{f \in \mathbb{R}^n} \mathcal{E}(f), \text{ where } \mathcal{E}(f) \overset{\text{def.}}{=} \frac{1}{2} \|f \star h - y\|^2 + \lambda \|Gf\|_{1,\varepsilon}$$

Compute $\nabla \mathcal{E}$.

[⋆⋆] **Wasserstein distances (exam 2019)** We recall what the Wasserstein-1 distance on a metric space X equipped with a distance d can be computed as

$$W_1(\alpha,\beta) = \sup_f \left\{ \int f \mathrm{d}(\alpha-\beta) \text{ s.t. } \mathrm{Lip}(f) \leq 1 \right\}$$

where the Lipschitz constant of a function $f : \mathcal{X} \to \mathbb{R}$ is $\mathrm{Lip}(f) \overset{\text{def.}}{=} \sup_{x\neq y} \frac{|f(x)-f(y)|}{d(x,y)}$.

1. For some random vector Z on some space $\mathcal{Z}$ and some function $\Phi : \mathcal{Z} \times \mathcal{F} \to \mathbb{R}$ show that

   $$\sup_{f\in\mathcal{F}} \mathbb{E}_Z(\Phi(Z,f)) \leq \mathbb{E}_Z(\sup_{f\in\mathcal{F}} \Phi(Z,f))$$

   where $\mathbb{E}_Z$ is the expectation with respect to Z.

2. We consider $2n$ independent random variables $X = (x_i)_{i=1}^n$ and $Y = (y_j)_{j=1}^n$, which are identically distributed with law respectively α and β. We denote $\hat{\alpha} \overset{\text{def.}}{=} \frac{1}{n}\sum_i \delta_{x_i}$ and $\hat{\beta} \overset{\text{def.}}{=} \frac{1}{n}\sum_j \delta_{y_j}$. Show that

   $$\mathbb{E}(W_1(\hat{\alpha},\hat{\beta})) \geq W_1(\alpha,\beta)$$

   where the expectation is computed with respect to (X,Y).

3. If $\varphi : \mathcal{X} \to \mathcal{Y}$ is a Lipschitz mapping between two metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$, show that

   $$W_1(\varphi_{\#}\alpha, \varphi_{\#}\beta) \leq \mathrm{Lip}(\varphi) W_1(\alpha,\beta)$$

   where $\varphi_{\#}$ is the push-forward operator and $\mathrm{Lip}(\varphi) \overset{\text{def.}}{=} \sup_{x\neq y} \frac{d_{\mathcal{Y}}(f(x),f(y))}{d_{\mathcal{X}}(x,y)}$.

4. We recall that the W_2 distance on $\mathbb{R}^d$ is defined as

   $$W_2(\alpha,\beta) \overset{\text{def.}}{=} \min_{\pi_1=\alpha,\pi_2=\beta} \int \|x-y\|^2 \mathrm{d}\pi(x,y).$$

   We denote $m_\alpha = \int \mathrm{d}\alpha$ the mean vector. We denote $\alpha_0 = T_{-m_\alpha\#}\alpha$ the centered distribution, which uses the pushforward by $T_{-m_\alpha} : x \mapsto x - m_\alpha$. Show that

   $$W_2(\alpha,\beta)^2 = W_2(\alpha_0,\beta_0)^2 + \|m_\alpha - m_\beta\|^2.$$

5. Show without using optimal transport that the discrete measures $\mu_n \frac{1}{n}\sum_k \delta_{k/n}$ converge weakly toward the uniform measure $\mathcal{U}$ on $[0,1]$.

6. Compute $W_1(\mu_n, \mathcal{U})$ and $W_2(\mu_n, \mathcal{U})$. Why does this solve the previous question?

Solutions

Coordinate descent

1. The gradient of f is $\nabla f(w) = Aw - b$, hence it is canceled wrt to coordinate i when $\langle a_i, w\rangle = b_i$, with a_i the i-th column of A. In turn, this is canceled w.r.t. w_i for $w_i = \frac{1}{A_{ii}}(b_i - \sum_{j\neq i} A_{ij} w_j)$.

2. Here, we are optimizing a 1-d quadratic function of the form

$$\psi(x) = \alpha x^2 + \beta x + \gamma$$

for which simple computations yield

$$\psi(x^*) - \psi(x) = -\alpha(x - x^*)^2$$

In our case, we have $\alpha = \frac{1}{2}A_{ii}$, so we get

$$f(w^{t+1}) - f(w^t) = -\frac{1}{2}A_{ii}(w_i^t - \frac{1}{A_{ii}}(b_i - \sum_{j\neq i} A_{ij} w_j^t))^2$$

which gives the advertised result. We then simply uper-bound A_{ii} by A_{max}.

3. Taking inequalities in the correct order, we get

$$\|Aw - b\|^2 = \sum_{j=1}^{n} (Aw^t - b)_j^2 \tag{1}$$
$$\leq p(Aw^t - b)_i^2 \tag{2}$$

since $(Aw^t - b)_i^2$ is maximal. We therefore obtain the proposed inequality.

4. We have

$$\|Aw - b\|^2 = \|A(w - w^*)\|^2 \tag{3}$$
$$\geq \sigma_{min}(A)\langle A(w - w^*), w - w^*\rangle \tag{4}$$
$$= 2\sigma_{min}(A)(f(w) - f(w^*)) \tag{5}$$

Plugging everything together we obtain

$$f(w^{t+1}) - f(w^t) \leq -\frac{\sigma_{min}(A)}{pA_{\max}}(f(w^t) - f(w^*)) \tag{6}$$

and unfolding that recursion gives

$$f(w^t) - f(w^*) \leq (1 - \frac{\sigma_{min}(A)}{pA_{\max}})^t (f(w^0) - f^*)$$

this is linear convergence. This should be compared to gradient descent convergence rate which is

$$f(w^t) - f(w^*) \leq (1 - \frac{\sigma_{min}(A)}{\sigma_{max}(A)})^t (f(w^0) - f^*)$$

We have $A_{max} \leq \sigma_{max}(A)$, hence $\frac{\sigma_{min}(A)}{\sigma_{max}(A)} \leq \frac{\sigma_{min}(A)}{A_{\max}}$. In practice, doing p iterations of CD is about as costly as one iteration of GD, hence the convergence of CD is faster, about the same when $A_{max} = \sigma_{max}(A)$, and better and better as A_{max} gets smaller than $\sigma_{\max}$.

Inverse problems

L2 regularization

1. Since the discrete Fourier basis is orthogonal, we can separate the problem into the following subproblems:
$$\arg\min_{\hat{f}_k} \frac{1}{2}|\hat{f}_k\hat{h}_k - \hat{y}_k|^2 + \frac{\lambda}{2}|\hat{f}_k|^2.$$
At optimality we have
$$\hat{f}_k|\hat{h}_k|^2 - \hat{y}_k\hat{h}_k^* + \lambda\hat{f}_k = 0,$$
So
$$\hat{f}_k = \frac{\hat{h}_k^*}{|\hat{h}_k|^2 + \lambda}\hat{y}_k.$$

2. When $\lambda = 0$, $\left\|\hat{f}_k - \hat{f}_{0,k}\right\| = \left|\frac{\hat{\varepsilon}_k}{\hat{h}_k}\right|$. The noise is thus amplified for small values of $\hat{h}_k$, which is the case in high-frequency since h is a low-pass filter.

Sobolev regularization

1. By change of variable,
$$\langle Gf, u\rangle = \sum_i (f_i - f_{i-1})u_i = \sum_i f_i u_i - \sum_i f_i u_{i+1} = \sum_i f_i(u_i - u_{i+1})$$
So $G^\top u = (u_i - u_{i+1})_i$.

2. $g = [1, 0, \ldots, 0, -1]$, $\tilde{g} = [1, -1, 0, \ldots, 0]$, and $\ell = [2, -1, 0, \ldots, 0, -1]$. One can notice $\tilde{g}_i = g_{-i}$ and $\ell_i = g_i + \tilde{g}_i$

3. $\hat{g}_k = 1 - e^{\frac{2ik\pi}{n}}$, $\hat{\tilde{g}}_k = 1 - e^{-\frac{2ik\pi}{n}} = \hat{g}_k^*$, and $\hat{\ell}_k = 2 - e^{-\frac{2ik\pi}{n}} - e^{\frac{2ik\pi}{n}} = 2i\sin(k\pi/n)(e^{-\frac{ik\pi}{n}} - e^{\frac{ik\pi}{n}}) = 4\sin^2(k\pi/n) = |\hat{g}_k|^2$

4. Since the discrete Fourier basis is orthogonal, we can separate the problem into the following subproblems:
$$\arg\min_{\hat{f}_k} \frac{1}{2}|\hat{f}_k\hat{h}_k - \hat{y}_k|^2 + \frac{\lambda}{2}|\hat{f}_k\hat{g}_k|^2.$$
At optimality we have
$$\hat{f}_k|\hat{h}_k|^2 - \hat{y}_k\hat{h}_k^* + \lambda\hat{f}_k\hat{\ell}_k = 0,$$
So
$$\hat{f}_k = \frac{\hat{h}_k^*}{|\hat{h}_k|^2 + \lambda\hat{\ell}_k}\hat{y}_k.$$

5. This penalizes the high frequencies

TV regularization

1. $\nabla \left\|\cdot\right\|_{1,\varepsilon}(f) = \left(\frac{f_n}{\left\|f_n\right\|_\varepsilon}\right)_n = \mathcal{N}_\varepsilon(f)$, the smooth normalization.

2. We use the chain rule and conclude $\nabla J_{\mathrm{TV}}^\varepsilon(f) = G^\top\mathcal{N}_\varepsilon(Gf)$

3. We take h symmetrical so $\nabla\mathcal{E}(f) = h \star (f \star h - y) - G^\top\mathcal{N}_\varepsilon(Gf)$