

## Problem Set 10 — Solutions (Coordinate Descent)

**Exercise 53** (Alternative analysis for gradient descent). *Let  $f$  be smooth with constant  $L$  in the classical sense, and satisfy the PL inequality (10.4). Let the problem  $\min_{\mathbf{x}} f(\mathbf{x})$  have a non-empty solution set  $\mathcal{X}^*$ . Prove that gradient descent with a stepsize of  $1/L$  has a global linear convergence rate*

$$f(\mathbf{x}_t) - f^* \leq \left(1 - \frac{\mu}{L}\right)^t (f(\mathbf{x}_0) - f^*).$$

**Solution:** We combine smoothness  $f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2$  with the gradient descent update  $\mathbf{x}_{t+1} := \mathbf{x}_t - \frac{1}{L} \nabla f(\mathbf{x}_t)$ .

We then obtain

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &= f(\mathbf{x}_t) - \frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|^2 \\ &\leq f(\mathbf{x}_t) - \frac{\mu}{L} (f(\mathbf{x}_t) - f^*) \end{aligned}$$

the last two lines being the PL inequality. If we subtract  $f^*$  and apply recursively, we obtain the claimed linear rate:

$$f(\mathbf{x}_t) - f^* \leq \left(1 - \frac{\mu}{L}\right)^t (f(\mathbf{x}_0) - f^*).$$

**Exercise 54** (Importance Sampling). *Consider random coordinate descent with selecting the  $i$ -th coordinate with probability proportional to the  $L_i$  value, where  $L_i$  is the individual smoothness constant for each coordinate  $i$  as in (10.5).*

*When using a stepsize of  $1/L_{i_t}$ , prove that we obtain the faster rate of*

$$\mathbb{E}[f(\mathbf{x}_t) - f^*] \leq \left(1 - \frac{\mu}{d\bar{L}}\right)^t [f(\mathbf{x}_0) - f^*],$$

where  $\bar{L} = \frac{1}{d} \sum_{i=1}^d L_i$  now is the average of all coordinate-wise smoothness constants. Note that this value can be much smaller than the global  $L$  we have used above, since that one was required to hold for all  $i$  so has to be chosen as  $L = \max_i L_i$  instead.

*Can you come up with an example from machine learning where  $\bar{L} \ll L$ ?*

**Solution:** We combine the coordinate descent update  $\mathbf{x}_{t+1} := \mathbf{x}_t - \frac{1}{L_{i_t}} \nabla_{i_t} f(\mathbf{x}_t) \mathbf{e}_{i_t}$ , with the property of coordinate-wise smoothness,  $f(\mathbf{x} + \gamma \mathbf{e}_i) \leq f(\mathbf{x}) + \gamma \nabla_i f(\mathbf{x}) + \frac{L_i}{2} \gamma^2$ , resulting in

$$\begin{aligned} f(\mathbf{x}_{t+1}) &= f(\mathbf{x}_t - \frac{1}{L_{i_t}} \nabla_{i_t} f(\mathbf{x}_t) \mathbf{e}_{i_t}) \\ &\leq f(\mathbf{x}_t) - \frac{1}{L_{i_t}} |\nabla_{i_t} f(\mathbf{x}_t)|^2 + \frac{L_{i_t}}{2} \left| \frac{1}{L_{i_t}} \nabla_{i_t} f(\mathbf{x}_t) \right|^2 \\ &= f(\mathbf{x}_t) - \frac{1}{2L_{i_t}} |\nabla_{i_t} f(\mathbf{x}_t)|^2. \end{aligned}$$

Taking the average we have:

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x}_{t+1})] &\leq f(\mathbf{x}_t) - \frac{1}{2} \mathbb{E}\left[\frac{1}{L_{i_t}} |\nabla_{i_t} f(\mathbf{x}_t)|^2\right] \\
&= f(\mathbf{x}_t) - \frac{1}{2} \sum_i \left[ \frac{L_i}{\sum_j L_j} \frac{1}{L_i} |\nabla_i f(\mathbf{x}_t)|^2 \right] \\
&= f(\mathbf{x}_t) - \frac{1}{2 \sum_j L_j} \|\nabla f(\mathbf{x}_t)\|^2
\end{aligned}$$

**Exercise 55.** Derive the solution to exact coordinate minimization for the Lasso problem (10.8), for the  $i$ -th coordinate. Write  $A_{-i}$  for the  $(d-1) \times n$  matrix obtained by removing the  $i$ -th column from  $A$ , and same for the vector  $\mathbf{x}_{-i}$  with one entry removed accordingly.

**Solution:** Use the subgradient optimality condition for unconstrained convex minimization, applied to the single coordinate problem.

$$2A_i^\top A_i x_i + 2A_i^\top (A_{-i} \mathbf{x}_{-i} - \mathbf{b}) + \lambda s$$

for  $s \in \partial|x_i|$  (subgradient of the univariate absolute value function).

$$\frac{\partial}{\partial x_i} [\|A\mathbf{x} - \mathbf{b}\|^2 + \lambda \|\mathbf{x}\|_1] = 2A_i^\top [A\mathbf{x} - \mathbf{b}] + \lambda s \text{ (with } s \in \partial|x_i|).$$

The previous partial derivative equals to zero, therefore:  $0 \stackrel{!}{=} 2A_i^\top A_i x_i + 2A_i^\top (A_{-i} \mathbf{x}_{-i} - \mathbf{b}) + \lambda s$  where  $A_i^\top$  is the  $i^{th}$  column of  $A$ .

This gives us:

$$\begin{aligned}
x_i &= \frac{-A_i^\top (A_{-i} \mathbf{x}_{-i} - \mathbf{b}) - \frac{1}{2} \lambda s}{A_i^\top A_i} \\
&= \frac{A_i^\top (\mathbf{b} - A_{-i} \mathbf{x}_{-i})}{\|A_i\|^2} - \frac{\lambda s}{2\|A_i\|^2} \\
&= S_{\frac{\lambda/2}{\|A_i\|^2}} \left( \frac{A_i^\top (\mathbf{b} - A_{-i} \mathbf{x}_{-i})}{\|A_i\|^2} \right)
\end{aligned}$$

The left  $S$  operator corresponds to soft thresholding, defined as  $S_a(b) = \begin{cases} 0, & |b| < a, \\ b - a & b > a, \\ b + a & b < -a \end{cases}$ .