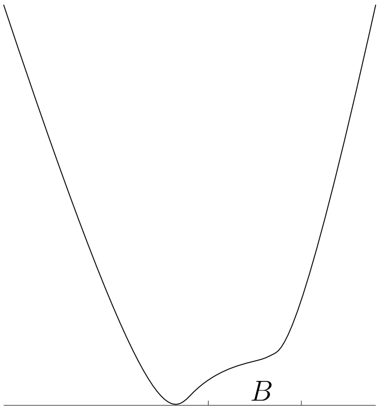
Labs
Optimization for Machine Learning
Spring 2025

EPFL
School of Computer and Communication Sciences
Nicolas Flammarion
github.com/epfml/OptML_course

Problem Set 8 — Solutions (Coordinate Descent)

Exercise 58. *Provide an example of a nonconvex function that satisfies the PL inequality 9.1!*

Solution: The key insight is the following: starting from any strongly convex function f, we can tweak f in any compact set B not containing the global minimum, without compromising the PL inequality. We can actually apply any tweaks that we want, as long as the function stays continuous and has no vanishing gradients in B. The reason is that over compact B, we always find a μ such that the PL inequality holds over B, as long as there are no vanishing gradients. Hence, we can for example start from $f(x) = x^2$ and introduce a non-convexity somewhere:



Exercise 59. *Prove Theorem 9.7! Can you come up with an example from machine learning where $\bar{L} \ll L = \max_{i=1}^d L_i$?*

Solution: This is very similar to the proof of Theorem 9.6. Sufficient decrease according to Lemma 9.5 yields

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L_i}|\nabla_i f(\mathbf{x}_t)|^2,$$

if coordinate i is chosen. By taking the expectation of both sides with respect to the choice of i, we have

$$\begin{aligned}
\mathbb{E}\left[f(\mathbf{x}_{t+1})|\mathbf{x}_t\right] &\leq f(\mathbf{x}_t) - \sum_{i=1}^d \frac{L_i}{\sum_{j=1}^d L_j}\frac{1}{2L_i}|\nabla_i f(\mathbf{x}_t)|^2 \\
&= f(\mathbf{x}_t) - \frac{1}{2\sum_{j=1}^d L_j}\sum_{i=1}^d |\nabla_i f(\mathbf{x}_t)|^2 \\
&= f(\mathbf{x}_t) - \frac{1}{2d\bar{L}}\|\nabla f(\mathbf{x}_t)\|^2 \\
&\leq f(\mathbf{x}_t) - \frac{\mu}{d\bar{L}}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \quad \text{(PL inequality (9.1))}.
\end{aligned}$$

Subtracting $f(\mathbf{x}^\star)$ from both sides, we therefore obtain

$$\mathbb{E}[f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star)|\mathbf{x}_t] \leq \left(1 - \frac{\mu}{d\bar{L}}\right)(f(\mathbf{x}_t) - f(\mathbf{x}^\star)).$$

Taking expectations (over $\mathbf{x}_t$), we obtain

$$\mathbb{E}[f(\mathbf{x}_{t+1}) - f(\mathbf{x}^\star)] \leq \left(1 - \frac{\mu}{d\bar{L}}\right)\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^\star)].$$

The statement follows.

Exercise 60. *Derive the solution to exact coordinate minimization for the Lasso problem (9.12), for the i-th coordinate. Write A_{-i} for the $n \times (d-1)$ matrix obtained by removing the i-th column from A, and same for the vector $\mathbf{x}_{-i}$ with one entry removed accordingly.*

Solution: We use the subgradient optimality condition for unconstrained convex minimization, applied to the single coordinate problem. A subgradient of this univariate objective can be written as

$$\begin{aligned}\frac{\partial}{\partial x_i}\left[\|A\mathbf{x}-\mathbf{b}\|^2+\lambda\|\mathbf{x}\|_1\right] &= 2A_i^\top[A\mathbf{x}-\mathbf{b}]+\lambda s \\ &= 2A_i^\top A_i x_i + 2A_i^\top(A_{-i}\mathbf{x}_{-i}-\mathbf{b})+\lambda s\end{aligned}$$

for $s \in \partial|x_i|$ being any subgradient of the (again univariate) absolute value function, and A_i is the i^{th} column of A.

At optimality, the previous partial derivative equals to zero, i.e. $0 \overset{!}{=} 2A_i^\top A_i x_i + 2A_i^\top(A_{-i}\mathbf{x}_{-i}-\mathbf{b})+\lambda s$.

Solving for x_i, this gives us:

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

The left S operator corresponds to soft thresholding, defined as

$$S_a(b) := \begin{cases} 0, & |b| \le a, \\ b-a & b > a, \\ b+a & b < -a \end{cases} .$$