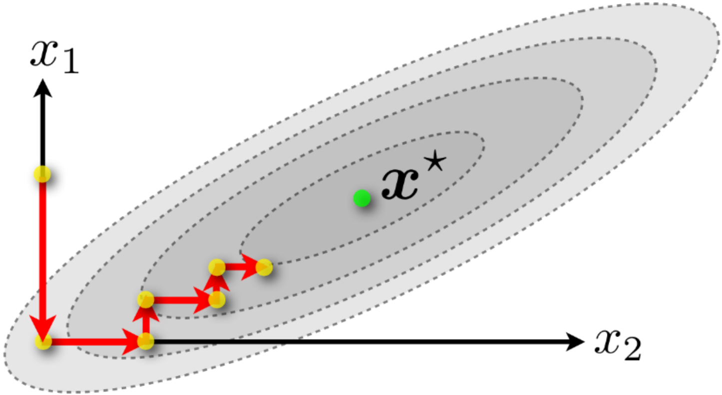
Coordinate descent

Axel Böhm

January 17, 2022

Coordinate Descent

Goal: Find $x^* \in \mathbb{R}^d$ minimizing $f(x)$.



Observation: Decrease in function value, but not in distance to solution.

Coordinate Descent

Modify only one coordinate per step:

$$\text{select } i_k \in \{1, \ldots, d\}$$
$$x_{k+1} = x_k + \gamma e_{i_k}$$

where e_i is the i-th unit basis vector. Two main variants:

- Gradient-based stepsize:

$$x_{k+1} = x_k - \frac{1}{L} \nabla_{i_k} f(x_k) e_{i_k}$$

- Exact coordinate minimization:
  Solve the **scalar** problem $\arg\min_{\gamma \in \mathbb{R}} f(x_k + \gamma e_{i_k})$.
  - *hyperparameter free*

Randomized Coordinate Descent

How to choose the coordinate?

$$\text{select } i_k \in \{1, \ldots, d\} \text{ uniformly at random}$$
$$x_{k+1} = x_k + \gamma e_{i_k}$$

- Faster convergence than gradient descent
  (if coordinate step is d times cheaper than full gradient step)

Technical assumptions

Coordinate-wise smoothness:

$$f(x+\gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L}{2}\gamma^2, \quad \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}, \forall i \in [d]$$

Is equivalent to coordinate-wise Lipschitz gradient:

$$|\nabla_i f(x+\gamma e_i) - \nabla_i f(x)| \leq L|\gamma|$$

- Additionally we assume strong convexity

Convergence: Linear rate

Theorem

Let f be coordinate-wise smooth with constant L and μ-strongly convex, then `randomized coordinate descent` *with stepsize* $1/L$

$$x_{k+1} = x_k - \frac{1}{L}\nabla_{i_k} f(x_k) e_{i_k}$$

where $i_k \sim Unif(1, \ldots, d)$

$$\mathbb{E}[f(x_k) - f^*] \leq \left(1 - \frac{\mu}{dL}\right)^k (f(x_0) - f^*)$$

Compare to rate of gradient descent.

Proof

By using smoothness we obtain

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla_{i_k} f(x_k)\|^2$$

Taking the expectation w.r.t. i

$$\begin{aligned}\mathbb{E}[f(x_{k+1})] &\leq f(x_k) - \frac{1}{2L}\mathbb{E}[|\nabla_{i_k} f(x_k)|^2] \\ &= f(x_k) - \frac{1}{2L}\frac{1}{d}\sum_i |\nabla_i f(x_k)|^2 \\ &= f(x_k) - \frac{1}{2dL}\|\nabla f(x_k)\|^2. \quad \square\end{aligned}$$

Lemma: Strong convexity implies **PL**: $\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*)$
Therefore, by subtracting f^* on both sides we get the statement of the theorem.

Polyak-Łojasiewicz (PL) Condition

Definition

f satisfies the PL condition if the following holds for some $\mu > 0$

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*).$$

Lemma

Strong convexity implies PL.

Proof Strong convexity gives

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2.$$

Minimizing each side w.r.t. y gives

$$f(x^*) \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2.$$

Linear convergence without strong convexity

PL is weaker than strong convexity (doesn't even imply convexity).

Examples satisfying PL

Let $f := g \circ A$ for strongly convex g and *arbitrary* matrix A, see **least squares regression**.

Corollary (Linear convergence for PL)

Same conditions as before but PL instead of strong convexity yields:

$$\mathbb{E}[f(x_k) - f^*] \leq \left(1 - \frac{\mu}{dL}\right)^k (f(x_0) - f^*)$$

Importance sampling

Uniform random selection is not always the best!

- Individual smoothness constants L_i for each coordinate i

$$f(x+\gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2$$

Coordinate descent with selection probabilities $P[i_k = i] = \frac{L_i}{\sum_i L_i}$ and stepsize $1/L_{i_k}$ converges with the faster rate

$$\mathbb{E}[f(x_k) - f^*] \leq \left(1 - \frac{\mu}{d\bar{L}}\right)^k (f(x_0) - f^*),$$

where $\bar{L} = \frac{1}{d}\sum_{i=1}^d L_i$.

Often $\bar{L} \ll L = \max_i L_i$!

Steepest Coordinate Descent

Selection rule given by

$$i_k = \arg\max_{i \in [d]} |\nabla_i f(x_k)|$$

"Greedy", Gauss-Southwell or **steepest** coordinate descent.
Drawback: requires computation of full gradient if you do not have additional knowledge.

Convergence of Steepest Coordinate Descent

Has same convergence rate as for random coordinate descent.
Use the fact that *max* is larger than *average*

$$\max_i |\nabla_i f(x)|^2 \geq \frac{1}{d} \sum_{i=1}^{d} |\nabla_i f(x)|^2,$$

Corollary

Steepest Coordinate Descent with stepsize $1/L$

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{dL}\right)^k (f(x_0) - f^*)$$

Benefit is not clear: more expensive iterations but same bound.

Faster Convergence of Steepest Coordinate Descent

Faster convergence when measuring strong convexity of f w.r.t 1-norm instead of the standard Euclidean norm, i.e.

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu_1}{2} \|x - y\|_1^2.$$

Theorem

Let f be coordinate-wise smooth with constant L and μ_1-strongly convex, w.r.t. the 1-norm. Then `steepest coordinate descent` *with stepsize* $1/L$

$$f(x_k) - f^* \leq \left(1 - \frac{\mu_1}{L}\right)^k (f(x_0) - f^*)$$

Compare this to previous contraction factor of $(1 - \frac{\mu}{dL})$.

We always have

$$\frac{\mu}{d} \leq \mu_1 \leq \mu.$$

Faster Convergence of Steepest Coordinate Descent II

Proof of previous theorem is same as before, but using the lemma

Lemma

Let f be μ_1-strongly convex with respect to the ℓ_1-norm, then

$$\frac{1}{2}\|\nabla f(x)\|_\infty^2 \geq \mu_1(f(x) - f^*)$$

Faster convergence on quadratics

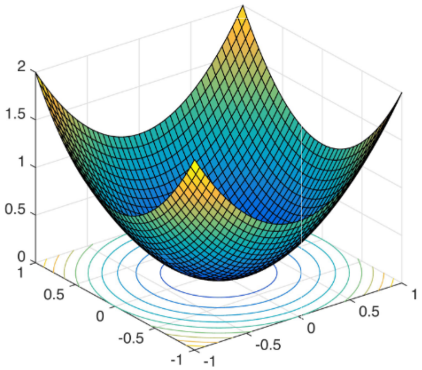
- If f is a quadratic with diagonal Hessian, we can show

$$\mu = \min_i \lambda_i \quad \text{and} \quad \mu_1 = \frac{1}{\sum_i \lambda_i}$$

- If all λ_i are equal:
  - No advantage to GS
- One very large λ_i
  - GS and random still similar
- One very small λ_i
  - GS bound can be much better $\mu_1 \approx \mu$

Non-smooth objectives

Proved everything for smooth f. What about nonsmooth?



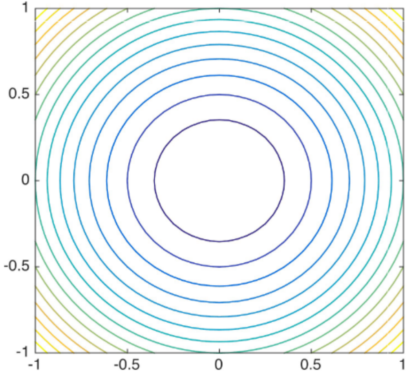
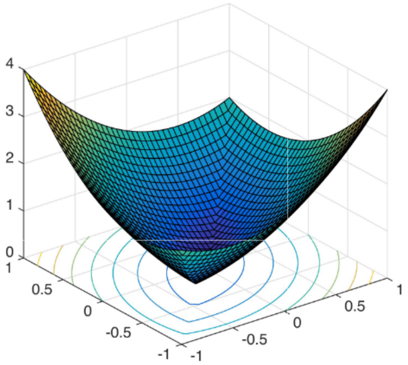


Figure: Example of a smooth function $f(x) = \|x\|^2$.

Non-smooth objectives

For general nonsmooth f coordinate descent fails and gets stuck



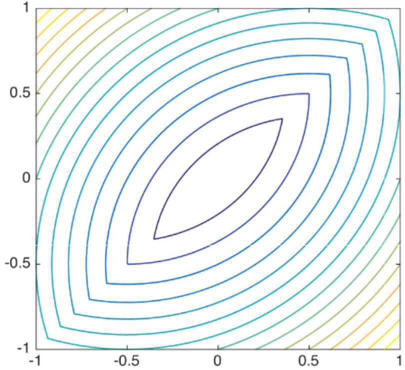
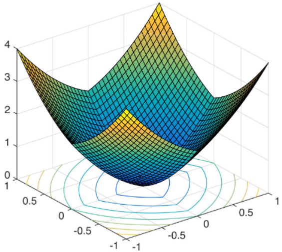


Figure: Example of a nonsmooth function $f(x) = \|x\|^2 + |x_1 - x_2|$.

Non-smooth separable objectives

If nonsmooth function is separable we can get convergence:

$$f(x) = g(x) + h(x) \quad \text{with} \quad h(x) = \sum_i h_i(x_i)$$



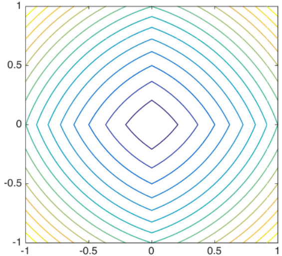


Figure: A nonsmooth but separable function $f(x) = \|x\|^2 + \|x\|_1$.

Randomized coordinate descent on non-strongly objectives

Theorem

Let f be coordinate-wise smooth with constant L and convex, then `randomized coordinate descent` *with stepsize $1/L$ yields*

$$\mathbb{E}[f(x_k) - f^*] \leq \frac{2Ld\|x_0 - x^*\|^2}{k}$$

same observation as in the strongly convex case.

Cyclic coordinate descent

Theorem

Let f be coordinate-wise smooth with constant L then `cyclic coordinate descent` *with stepsize $1/L$ achieves for*

- *convex objective*

$$\mathbb{E}[f(x_k) - f^*] \leq \frac{4L(d+1)\|x_0 - x^*\|^2}{k}$$

- *and for μ-strongly convex objectives*

$$\mathbb{E}[f(x_k) - f^*] \leq \left(1 - \frac{\mu}{2(d+1)L}\right)^k (f(x_0) - f^*)$$

Again, randomized version was better.

Some more thoughts

- minimize all coordinates individually (**in parallel**)
- can use blocks of coordinates instead of individual ones

State of the art for generalized linear models $f(x) := g(Ax) + \sum_i^d h_i(x)$

- Regression, classification (with regularizers)