

# Randomized Coordinate Gradient Descent

Pontus Giselsson

# Outline

- **Coordinate proximal gradient method**
- Coordinate-wise smoothness
- A fundamental inequality
- Nonconvex setting
- Convex setting
- Strongly convex setting
- Rate comparison to proximal gradient method

# Composite problem form

- Consider composite problems of the form

$$\underset{x}{\text{minimize}} f(x) + \underbrace{\sum_{i=1}^n g_i(x_i)}_{g(x)}$$

where

- $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is smooth (will be refined)
- $g : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  is closed convex and separable
- Problem structure includes:
  - Training problems with  $\|x\|_1$  or  $\|x\|_2^2$  regularization
  - Dual SVM problem formulation

## Coordinate proximal gradient descent

- Compute proximal gradient step, update random coordinate  $j$ :

$j \in \{1, \dots, n\}$  is randomly chosen with uniform probability

$$x_j^{k+1} = \text{prox}_{\gamma_j g_j}(x_j^k - \gamma_j \nabla f(x^k)_j)$$
$$x_i^{k+1} = x_i^k \text{ for all } i \neq j$$

- Comments:
  - We use super-scripts for iteration and sub-script for coordinate
  - Can take blocks of coordinates (will treat single-coordinate case)
  - Algorithm analysis very similar to proximal gradient descent
  - Individual step-size  $\gamma_j$  for every coordinate

## Coordinate proximal gradient descent – Reformulation

- Let  $\Gamma := \text{diag}(\gamma_1, \dots, \gamma_n)$ , then we can write the  $x_j$  update as

$$x_j^{k+1} = (\text{prox}_g^{\Gamma^{-1}}(x^k - \Gamma \nabla f(x^k)))_j$$

where  $\text{prox}_g^H(z) := \text{argmin}_x (g(x) + \frac{1}{2} \|x - z\|_H^2)$

- This holds since  $\Gamma$  is diagonal,  $g$  and  $\|\cdot\|_{\Gamma^{-1}}$  are separable:

$$\begin{aligned} & \text{prox}_g^{\Gamma^{-1}}(x^k - \Gamma \nabla f(x^k)) \\ &= \text{argmin}_x (g(x) + \frac{1}{2} \|x - (x^k - \Gamma \nabla f(x^k))\|_{\Gamma^{-1}}^2) \\ &= \text{argmin}_x \left( \sum_{i=1}^n g_i(x_i) + \frac{1}{2\gamma_i} (x_i - (x_i^k - \gamma_i \nabla f(x^k)_i))^2 \right) \end{aligned}$$

where optimal  $x_j$  is found by optimizing only  $j$ th part of the sum

- Updates one coordinate of full scaled proximal gradient step

## Efficient evaluation

- The core update is

$$x_j^{k+1} = \text{prox}_{\gamma_j g_j}(x_j^k - \gamma_j \nabla f(x^k)_j)$$

- Assume update cost roughly  $\frac{1}{n}$  compared to full proximal gradient
  - Then  $n$  coordinate updates at same cost as one full update
  - In this scenario, coordinate gradient descent often faster
- Computational cost of  $\text{prox}_{\gamma_j g_j}$ 
  - 1D optimization problem
  - Often closed form solution or fast to evaluate
  - Performed at cost  $\frac{1}{n}$  compared to full prox due to separability of  $g$
- Computational cost of  $\nabla f(x^k)_j$  – element  $j$  of full gradient
  - This is often the costly part of the algorithm
  - Requires in general to compute full gradient, then pick element
  - Method efficient if cost roughly  $\frac{1}{n}$  of full gradient cost

## Efficient coordinate gradient evaluation – Quadratics

- Let  $f(x) = \frac{1}{2}x^T Px + q^T x$  with  $P \in \mathbb{R}^{n \times n}$ , then:

$$\nabla f(x)_j = (Px)_j + q_j = P_j^T x + q_j$$

where  $P_j \in \mathbb{R}^n$  is  $j$ th column of  $P$

- Uses one of  $n$  columns in  $P$  and one of  $n$  elements in  $q$
- Coordinate gradient evaluated at cost  $\frac{1}{n}$  of full gradient

## Efficient coordinate gradient evaluation

- Let  $\nabla f(x) = L^T(\sigma(Lx) - b)$  with
  - matrix  $L \in \mathbb{R}^{m \times n}$ ,  $L_j \in \mathbb{R}^m$  is  $j$ th column in  $L$ , vector  $b \in \mathbb{R}^m$
  - maximal monotone mapping  $\sigma : \mathbb{R}^m \rightarrow \mathbb{R}^m$

then  $\nabla f(x)$  is maximally monotone and  $f$  convex

- Coordinate gradient

$$\nabla f(x)_j = (L^T(\sigma(Lx) - b))_j = L_j^T(\sigma(Lx) - b)$$

- Assume we know  $z = Ly$  at point  $y = (x_1, \dots, y_l, \dots, x_n)$ :

$$Lx = Ly + L(x - y) = z + L_l(x_l - y_l)$$

where  $x_l - y_l$  is a scalar, and coordinate gradient

$$\nabla f(x)_j = L_j^T(\sigma(z + L_l(x_l - y_l)) - b)$$

can be updated at roughly  $\frac{1}{n}$  of cost for a full gradient



## Proximal gradient method – Convergence rates

- We will analyze coordinate method in different settings:
  - Nonconvex
    - $O(1/k)$  convergence for squared residual
  - Convex
    - $O(1/k)$  convergence for function values
  - Strongly convex
    - Linear convergence in distance to solution
- First two rates based on a *fundamental inequality* for the method
- Same rates as for proximal gradient, but improved constants

# Outline

- Coordinate proximal gradient method
- **Coordinate-wise smoothness**
- A fundamental inequality
- Nonconvex setting
- Convex setting
- Strongly convex setting
- Rate comparison to proximal gradient method

## Coordinate-wise smoothness

- For proximal gradient method we assume quadratic upper bound
- This is implied, for instance, by smoothness of  $f$
- In coordinate method, we will exploit *coordinate-wise smoothness*

## Coordinate-wise smoothness – Definition

- Coordinate-wise  $\beta_j$ -Lipschitz continuity, let  $y_i = x_i$  for all  $i \neq j$

$$|\nabla f(x)_j - \nabla f(y)_j| \leq \beta_j |x_j - y_j|$$

- Similar to for smoothness, this is equivalent to that

$$f(y) \leq f(x) + \nabla f(x)_j (y_j - x_j) + \frac{\beta_j}{2} (x_j - y_j)^2$$

$$f(y) \geq f(x) + \nabla f(x)_j (y_j - x_j) - \frac{\beta_j}{2} (x_j - y_j)^2$$

for all  $x$  and  $y$  such that  $y_i = x_i$  for all  $i \neq j$

- We can explicitly express coordinate with  $y = x + te_j$

$$f(x + te_j) \leq f(x) + \nabla f(x)_j t + \frac{\beta_j}{2} t^2$$

$$f(x + te_j) \geq f(x) + \nabla f(x)_j t - \frac{\beta_j}{2} t^2$$

where  $e_j$  is  $j$ th standard basis vector in  $\mathbb{R}^n$

- We will assume that such  $\beta_j$  exist

## Coordinate descent – Interpretation

- In proximal gradient,  $f$  replaced by smoothness upper bound
- In coordinate gradient, replace by coordinate-smoothness:

$$\begin{aligned}x_j^{k+1} &= \operatorname{argmin}_{y_i} (f(x^k) + \nabla f(x^k)_j (y_j - x_j^k) + \frac{1}{2\gamma_j} (y_j - x_j^k)^2 + g_j(y_j)) \\ &= \operatorname{argmin}_{y_i} (g_j(y_j) + \frac{1}{2\gamma_j} (y_j - (x_j^k - \gamma_j \nabla f(x^k)_j))^2) \\ &= \operatorname{prox}_{\gamma_j g_j} (x_j^k - \gamma_j \nabla f(x^k)_j)\end{aligned}$$

which is the  $j$ th component update

## Comparison to smoothness

- By  $\beta$ -smoothness of  $f$  we have for all  $x, y \in \mathbb{R}^n$ :

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{\beta}{2}\|x - y\|_2^2$$

- If we restrict  $y$  and  $x$  so that  $y_i = x_i$  for all  $i \neq j$  then

$$f(y) \leq f(x) + \nabla f(x)_j(y_j - x_j) + \frac{\beta}{2}(x_j - y_j)^2$$

- So  $\beta$  is coordinate-wise smoothness constant, we have for all  $j$ :

$$\beta_j \leq \beta$$

## Coordinate smoothness for quadratics

- Suppose that  $f(x) = \frac{1}{2}x^T Px + q^T x$  is a convex quadratic
- Then  $f$  is  $p_{jj}$ -coordinate-wise smooth, let  $y = x + te_j$ , then

$$\begin{aligned}f(x + te_j) &= \frac{1}{2}(x + te_j)^T P(x + te_j) + q^T(x + te_j) \\&= \frac{1}{2}x^T Px + q^T x + (Px)^T(te_j) + q^T te_j + \frac{1}{2}t^2 e_j^T P e_j \\&= \frac{1}{2}x^T Px + q^T x + (Px + q)_j t + \frac{p_{jj}}{2} t^2 \\&= f(x) + \nabla f(x)_j t + \frac{p_{jj}}{2} t^2\end{aligned}$$

which proves the claim

- Note that we have equality, which also implies

$$f(y) = f(x) + \nabla f(x)_j (y_j - x_j) + \frac{p_{jj}}{2} (y_j - x_j)^2$$

for all  $y$  and  $x$  such that  $y_i = x_i$  for  $i \neq j$

## Coordinate descent for quadratics

- Let  $f(x) = \frac{1}{2}x^T Px + q^T x$  and use  $\gamma_j = \frac{1}{p_{jj}}$  in algorithm
- The coordinate descent method becomes, with  $y = x^k + te_j$ :

$$\begin{aligned}x_j^{k+1} &= \operatorname{argmin}_{y_j} (f(x^k) + \nabla f(x^k)_j (y_j - x_j^k) + \frac{p_{jj}}{2} (y_j - x_j^k)^2 + g_j(y_j)) \\&= \operatorname{argmin}_t (f(x^k) + \nabla f(x^k)_j t + \frac{p_{jj}}{2} t^2 + g_j(x_j^k + t)) \\&= \operatorname{argmin}_t (f(x^k + te_j) + g_j(x_j^k + t)) \\&= \operatorname{argmin}_t (f(x^k + te_j) + g(x^k + te_j))\end{aligned}$$

- This choice of  $\gamma_j$  gives here coordinate-wise minimization

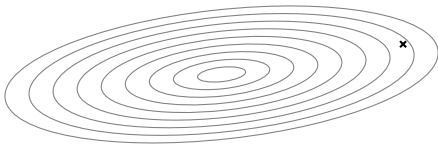


## Example – Uniform smoothness constant

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-sizes  $\gamma_1 = \gamma_2 = \frac{1}{\beta}$

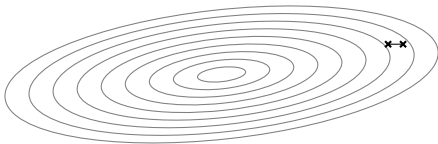


## Example – Uniform smoothness constant

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-sizes  $\gamma_1 = \gamma_2 = \frac{1}{\beta}$

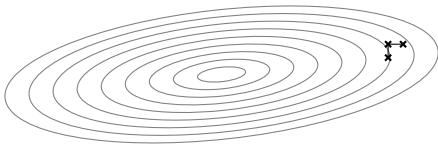


## Example – Uniform smoothness constant

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-sizes  $\gamma_1 = \gamma_2 = \frac{1}{\beta}$

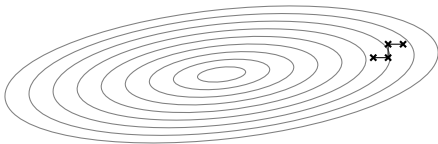


## Example – Uniform smoothness constant

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-sizes  $\gamma_1 = \gamma_2 = \frac{1}{\beta}$

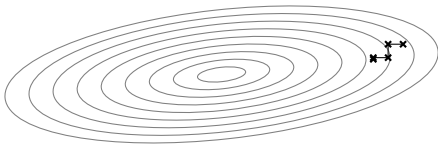


## Example – Uniform smoothness constant

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-sizes  $\gamma_1 = \gamma_2 = \frac{1}{\beta}$

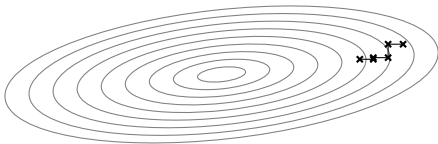


## Example – Uniform smoothness constant

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-sizes  $\gamma_1 = \gamma_2 = \frac{1}{\beta}$

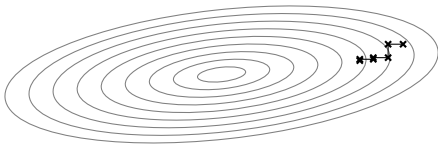


## Example – Uniform smoothness constant

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-sizes  $\gamma_1 = \gamma_2 = \frac{1}{\beta}$

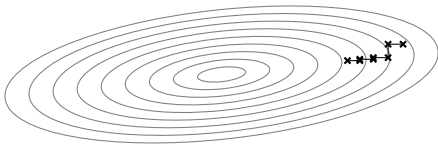


## Example – Uniform smoothness constant

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-sizes  $\gamma_1 = \gamma_2 = \frac{1}{\beta}$



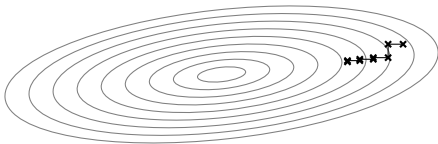


## Example – Uniform smoothness constant

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-sizes  $\gamma_1 = \gamma_2 = \frac{1}{\beta}$

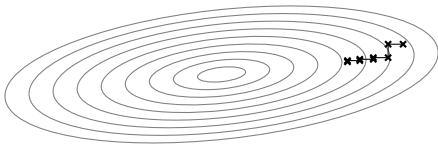


## Example – Uniform smoothness constant

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-sizes  $\gamma_1 = \gamma_2 = \frac{1}{\beta}$

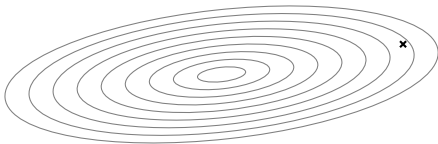


## Example – Individual smoothness constants

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size  $\gamma_1 = p_{11}^{-1} = 10$  and  $\gamma_2 = p_{22}^{-1} = 1$

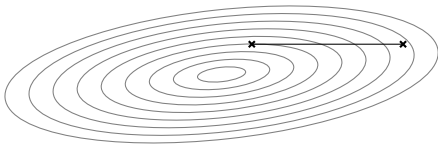


## Example – Individual smoothness constants

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size  $\gamma_1 = p_{11}^{-1} = 10$  and  $\gamma_2 = p_{22}^{-1} = 1$

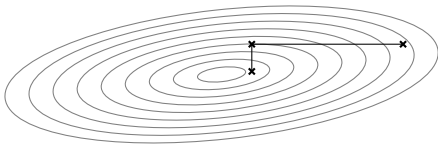


## Example – Individual smoothness constants

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size  $\gamma_1 = p_{11}^{-1} = 10$  and  $\gamma_2 = p_{22}^{-1} = 1$

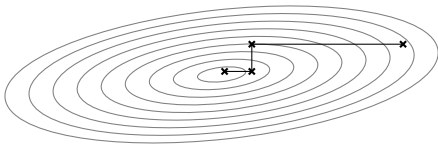


## Example – Individual smoothness constants

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size  $\gamma_1 = p_{11}^{-1} = 10$  and  $\gamma_2 = p_{22}^{-1} = 1$

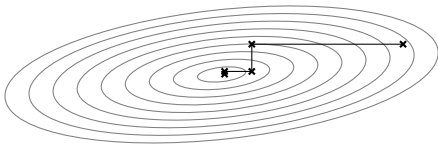


## Example – Individual smoothness constants

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size  $\gamma_1 = p_{11}^{-1} = 10$  and  $\gamma_2 = p_{22}^{-1} = 1$

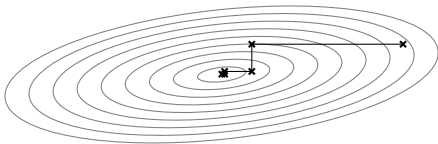


## Example – Individual smoothness constants

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size  $\gamma_1 = p_{11}^{-1} = 10$  and  $\gamma_2 = p_{22}^{-1} = 1$



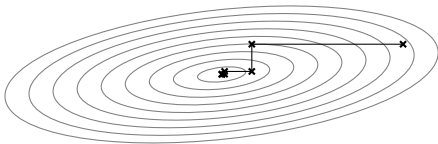


## Example – Individual smoothness constants

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size  $\gamma_1 = p_{11}^{-1} = 10$  and  $\gamma_2 = p_{22}^{-1} = 1$

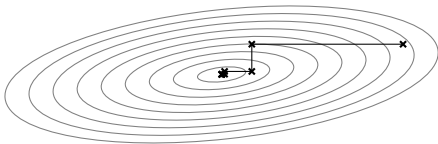


## Example – Individual smoothness constants

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size  $\gamma_1 = p_{11}^{-1} = 10$  and  $\gamma_2 = p_{22}^{-1} = 1$

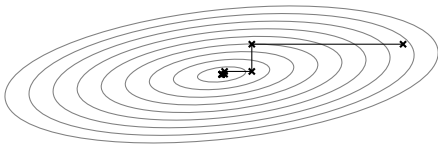


## Example – Individual smoothness constants

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size  $\gamma_1 = p_{11}^{-1} = 10$  and  $\gamma_2 = p_{22}^{-1} = 1$

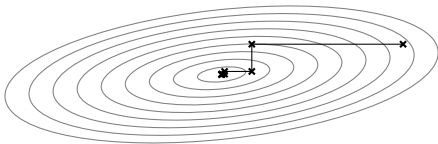


## Example – Individual smoothness constants

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size  $\gamma_1 = p_{11}^{-1} = 10$  and  $\gamma_2 = p_{22}^{-1} = 1$



# Outline

- Coordinate proximal gradient method
- Coordinate-wise smoothness
- **A fundamental inequality**
- Nonconvex setting
- Convex setting
- Strongly convex setting
- Rate comparison to proximal gradient method

## Coordinate proximal gradient descent

- Consider separable composite problems of the form

$$\underset{x}{\text{minimize}} \quad f(x) + \underbrace{\sum_{i=1}^n g_i(x_i)}_{g(x)}$$

- Will analyze coordinate proximal gradient method:

$j \in \{1, \dots, n\}$  is randomly chosen with uniform probability

$$x_j^{k+1} = \text{prox}_{\gamma_j g_j}(x_j^k - \gamma_j \nabla f(x^k)_j)$$
$$x_i^{k+1} = x_i^k \text{ for all } i \neq j$$

## Assumptions for fundamental inequality

- (i)  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is continuously differentiable (not necessarily convex)  
(ii)  $f$  is  $\beta_j$ -coordinate smooth, i.e., we have

$$f(y) \leq f(x) + \nabla f(x)_j (y_j - x_j) + \frac{\beta_j}{2} (x_j - y_j)^2$$

for all  $x, y \in \mathbb{R}^n$  such that  $y_i = x_i$  for all  $i \neq j$

- (iii)  $g : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  closed convex and separable  
(iv) A minimizer  $x^*$  exists and  $p^* = f(x^*) + g(x^*)$  is optimal value  
(v) Algorithm parameters  $\gamma_j > 0$

- Similar assumptions as for proximal gradient method
- Also results and proofs similar, but a bit more technical

## A fundamental inequality

For all  $z \in \mathbb{R}^n$ , the coordinate proximal gradient method satisfies

$$\begin{aligned} & \mathbb{E}[f(x^{k+1}) + g(x^{k+1})|x^k] \\ & \leq f(x^k) + \frac{1}{n}g(z) + \frac{1}{n}\nabla f(x^k)^T(z - x^k) + \frac{n-1}{n}g(x^k) \\ & \quad + \frac{1}{2}\mathbb{E}[(\beta_j - \gamma_j^{-1})(x_j^{k+1} - x_j^k)^2|x^k] \\ & \quad + \frac{1}{2}(\mathbb{E}[\gamma_j^{-1}(x_j^k - z_j)^2|x^k] - \mathbb{E}[\gamma_j^{-1}(x_j^{k+1} - z_j)^2|x^k]) \end{aligned}$$



## A fundamental inequality - Proof (1/3)

Using

- (a)  $\beta_j$ -coordinate smoothness of  $f$ , i.e., Assumption (ii)  
(b) Prox optimality condition: There exists  $s_j^{k+1} \in \partial g_j(x_j^{k+1})$

$$0 = s_j^{k+1} + \gamma_j^{-1}(x_j^{k+1} - (x_j^k - \gamma_j \nabla f(x^k)_j))$$

- (c) Subgradient:  $\forall z_j, g_j: g_j(z_j) \geq g_j(x_j^{k+1}) + s_j^{k+1}(z_j - x_j^{k+1})$

$$f(x^{k+1}) + g_j(x_j^{k+1})$$

$$(a) \leq f(x^k) + \nabla f(x^k)_j(x_j^{k+1} - x_j^k) + \frac{\beta_j}{2}(x_j^{k+1} - x_j^k)^2 + g_j(x_j^{k+1})$$

$$(c) \leq f(x^k) + \nabla f(x^k)_j(x_j^{k+1} - x_j^k) + \frac{\beta_j}{2}(x_j^{k+1} - x_j^k)^2$$

$$+ g_j(z_j) - s_j^{k+1}(z_j - x_j^{k+1})$$

$$(b) = f(x^k) + \nabla f(x^k)_j(x_j^{k+1} - x_j^k) + \frac{\beta_j}{2}(x_j^{k+1} - x_j^k)^2$$

$$+ g_j(z_j) + \gamma_j^{-1}(x_j^{k+1} - (x_j^k - \gamma_j \nabla f(x^k)_j))(z_j - x_j^{k+1})$$

$$= f(x^k) + g_j(z_j) + \nabla f(x^k)_j(z_j - x_j^k) + \frac{\beta_j}{2}(x_j^{k+1} - x_j^k)^2$$

$$+ \gamma_j^{-1}(x_j^{k+1} - x_j^k)(z_j - x_j^{k+1})$$

## A fundamental inequality – Proof (2/3)

- Now, let us use the equality

$$(x_j^{k+1} - x_j^k)(z_j - x_j^{k+1}) = \frac{1}{2}((x_j^k - z_j)^2 - (x_j^{k+1} - z_j)^2 - (x_j^k - x_j^{k+1})^2)$$

- Applying to previous inequality gives

$$\begin{aligned} & f(x^{k+1}) + g_j(x_j^{k+1}) \\ & \leq f(x^k) + g_j(z_j) + \nabla f(x^k)_j(z_j - x_j^k) + \frac{\beta_j}{2}(x_j^{k+1} - x_j^k)^2 \\ & \quad + \gamma_j^{-1}(x_j^{k+1} - x_j^k)(z_j - x_j^{k+1}) \\ & = f(x^k) + g_j(z_j) + \nabla f(x^k)_j(z_j - x_j^k) + \frac{\beta_j}{2}(x_j^{k+1} - x_j^k)^2 \\ & \quad + \frac{1}{2\gamma_j}((x_j^k - z_j)^2 - (x_j^{k+1} - z_j)^2 - (x_j^k - x_j^{k+1})^2) \\ & = f(x^k) + g_j(z_j) + \nabla f(x^k)_j(z_j - x_j^k) + \frac{\beta_j - \gamma_j^{-1}}{2}(x_j^{k+1} - x_j^k)^2 \\ & \quad + \frac{1}{2\gamma_j}((x_j^k - z_j)^2 - (x_j^{k+1} - z_j)^2) \end{aligned}$$

## A fundamental inequality – Proof (3/3)

- Now, take expected value conditioned on  $x^k$ :

$$\begin{aligned}\mathbb{E}[f(x^{k+1}) + g(x^{k+1})|x^k] &= \mathbb{E}[f(x^{k+1}) + g_j(x_j^{k+1}) + \sum_{i \neq j} g_i(x_i^k)|x^k] \\ &\leq \mathbb{E}[f(x^k) + g_j(z_j) + \nabla f(x^k)_j(z_j - x_j^k) + \frac{\beta_j - \gamma_j^{-1}}{2}(x_j^{k+1} - x_j^k)^2 \\ &\quad + \frac{1}{2\gamma_j}((x_j^k - z_j)^2 - (x_j^{k+1} - z_j)^2)|x^k] + \frac{n-1}{n} \sum_{i=1}^n g_i(x_i^k) \\ &= f(x^k) + \frac{1}{n}g(z) + \frac{1}{n}\nabla f(x^k)^T(z - x^k) \\ &\quad + \frac{1}{2}\mathbb{E}[(\beta_j - \gamma_j^{-1})(x_j^{k+1} - x_j^k)^2|x^k] + \frac{n-1}{n}g(x^k) \\ &\quad + \frac{1}{2}(\mathbb{E}[\gamma_j^{-1}(x_j^k - z_j)^2|x^k] - \mathbb{E}[\gamma_j^{-1}(x_j^{k+1} - z_j)^2|x^k])\end{aligned}$$

- This is the *fundamental inequality* that we wanted to prove

# Outline

- Coordinate proximal gradient method
- Coordinate-wise smoothness
- A fundamental inequality
- **Nonconvex setting**
- Convex setting
- Strongly convex setting
- Rate comparison to proximal gradient method

## Nonconvex setting

- We will analyze the coordinate proximal gradient method

$$j \in \{1, \dots, n\} \text{ is randomly chosen with uniform probability}$$
$$x_j^{k+1} = \text{prox}_{\gamma_j g_j}(x_j^k - \gamma_j \nabla f(x^k)_j)$$
$$x_i^{k+1} = x_i^k \text{ for all } i \neq j$$

in a nonconvex setting for solving

$$\underset{x}{\text{minimize}} f(x) + \underbrace{\sum_{i=1}^n g_i(x_i)}_{g(x)}$$

- Will show sublinear convergence
- Analysis based on *A fundamental inequality*

## Nonconvex setting – Assumptions

- (i)  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is continuously differentiable (not necessarily convex)  
(ii)  $f$  is  $\beta_j$ -coordinate smooth, i.e., we have

$$f(y) \leq f(x) + \nabla f(x)_j (y_j - x_j) + \frac{\beta_j}{2} (x_j - y_j)^2$$

for all  $x, y \in \mathbb{R}^n$  such that  $y_i = x_i$  for all  $i \neq j$

- (iii)  $g : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  closed convex and separable  
(iv) A minimizer  $x^*$  exists and  $p^* = f(x^*) + g(x^*)$  is optimal value  
(v) Algorithm parameters  $\gamma_j \in (0, \frac{2}{\beta_j})$

- Same as for fundamental inequality but restricted step-sizes

## Nonconvex setting – Analysis

- Use fundamental inequality

$$\begin{aligned} & \mathbb{E}[f(x^{k+1}) + g(x^{k+1})|x^k] \\ & \leq f(x^k) + \frac{1}{n}g(z) + \frac{1}{n}\nabla f(x^k)^T(z - x^k) + \frac{n-1}{n}g(x^k) \\ & \quad + \frac{1}{2}\mathbb{E}[(\beta_j - \gamma_j^{-1})(x_j^{k+1} - x_j^k)^2|x^k] \\ & \quad + \frac{1}{2}(\mathbb{E}[\gamma_j^{-1}(x_j^k - z_j)^2|x^k] - \mathbb{E}[\gamma_j^{-1}(x_j^{k+1} - z_j)^2|x^k]) \end{aligned}$$

- Set  $z = x^k$  to get

$$\begin{aligned} \mathbb{E}[f(x^{k+1}) + g(x^{k+1})|x^k] & \leq f(x^k) + g(x^k) \\ & \quad - \frac{1}{2}\mathbb{E}[(\frac{2}{\gamma_j} - \beta_j)(x_j^{k+1} - x_j^k)^2|x^k] \end{aligned}$$

## Expected value of residual

- Let  $B = \mathbf{diag}(\beta_1, \dots, \beta_n)$  and recall  $\Gamma = \mathbf{diag}(\gamma_1, \dots, \gamma_n)$
- The expected value of the residual satisfies

$$\begin{aligned} & \mathbb{E}\left[\left(\frac{2}{\gamma_j} - \beta_j\right)(x_j^{k+1} - x_j^k)^2 \mid x^k\right] \\ &= \frac{1}{n} \sum_{i=1}^n \left(\frac{2}{\gamma_i} - \beta_i\right) (\text{prox}_{\gamma_i g_i}(x_i^k - \gamma_i \nabla f(x^k)_i) - x_i^k)^2 \\ &= \frac{1}{n} \sum_{i=1}^n \left(\frac{2}{\gamma_i} - \beta_i\right) (\text{prox}_g^{\Gamma^{-1}}(x^k - \Gamma \nabla f(x^k)) - x^k)_i^2 \\ &= \frac{1}{n} \|\text{prox}_g^{\Gamma^{-1}}(x^k - \Gamma \nabla f(x^k)) - x^k\|_{2\Gamma^{-1}-B}^2 \end{aligned}$$



## Step-size requirement

- Fundamental inequality with  $z = x^k$  and previous expected value:

$$\mathbb{E}[f(x^{k+1}) + g(x^{k+1})|x^k] \leq f(x^k) + g(x^k) - \frac{1}{2n} \|\text{prox}_g^{\Gamma^{-1}}(x^k - \Gamma \nabla f(x^k)) - x^k\|_{2\Gamma^{-1}-B}^2$$

- The step-size requirement  $\gamma_j \in (0, \frac{2}{\beta_j})$  implies  $2\Gamma^{-1} - B \succ 0$
- Subtract  $p^*$ , take expectation, use law of total expectation:

$$\underbrace{\mathbb{E}[f(x^{k+1}) + g(x^{k+1}) - p^*]}_{V_{k+1}} \leq \underbrace{\mathbb{E}[f(x^k) + g(x^k) - p^*]}_{V_k} - \underbrace{\mathbb{E}[\frac{1}{2n} \|\text{prox}_g^{\Gamma^{-1}}(x^k - \Gamma \nabla f(x^k)) - x^k\|_{2\Gamma^{-1}-B}^2]}_{R_k}$$

where the bounds on the step-sizes make  $R_k$  nonnegative

## Lyapunov inequality consequences

- We showed Lyapunov inequality  $V_{k+1} \leq V_k - R_k$  with quantities

$$V_k = \mathbb{E}[f(x^k) + g(x^k) - p^*]$$

$$R_k = \mathbb{E}[\frac{1}{2n} \|\text{prox}_g^{\Gamma^{-1}}(x^k - \Gamma \nabla f(x^k)) - x^k\|_{2\Gamma^{-1}-B}^2]$$

- Consequences (similar to for proximal gradient method):
  - Expected function value is decreasing (may not go to  $p^*$ )
  - Expected residual is summable, since  $2\Gamma^{-1} - B \succ 0$ :

$$\sum_{l=0}^{\infty} \mathbb{E}[\|\text{prox}_g^{\Gamma^{-1}}(x^l - \Gamma \nabla f(x^l)) - x^l\|_2] < \infty$$

and residual converges almost surely to 0

- Expected value of best residual squared converges as  $O(1/k)$ :

$$\mathbb{E}[\min_{l=\{0, \dots, k\}} \|\text{prox}_g^{\Gamma^{-1}}(x^l - \Gamma \nabla f(x^l)) - x^l\|_{2\Gamma^{-1}-B}^2] \leq \frac{2n(f(x^0) + g(x^0) - p^*)}{k+1}$$

where Jensen's inequality used to swap  $\mathbb{E}$  and  $\min_l$

## Expected fixed-point residual convergence

What does  $\mathbb{E}[\|\text{prox}_g^{\Gamma^{-1}}(x^k - \Gamma \nabla f(x^k)) - x^k\|_2] \rightarrow 0$  imply?

- Since expected residual is nonnegative and summable

$$\|\text{prox}_g^{\Gamma^{-1}}(x^k - \Gamma \nabla f(x^k)) - x^k\|_2 \rightarrow 0$$

a.s., meaning algorithm realizations satisfy this with probability 1

- Let  $v^k = \text{prox}_g^{\Gamma^{-1}}(x^k - \Gamma \nabla f(x^k))$ , then

$$\partial g(v^k) + \nabla f(v^k) \ni \Gamma^{-1}(x^k - v^k) + \nabla f(v^k) - \nabla f(x^k) \rightarrow 0$$

- So:
  - $v^k$  sequence satisfies fixed-point characterization in limit
  - $x^k$  is arbitrarily close to  $v^k$
  - if  $x^k$  (sub)sequence converges to  $\bar{x}$ , so does  $v_k$ , and we have

$$\partial g(\bar{x}) + \nabla f(\bar{x}) \ni 0$$

(by closedness of graphs of maximal monotone operators)

# Outline

- Coordinate proximal gradient method
- Coordinate-wise smoothness
- A fundamental inequality
- Nonconvex setting
- **Convex setting**
- Strongly convex setting
- Rate comparison to proximal gradient method

## Convex setting

- We will analyze the coordinate proximal gradient method

$$j \in \{1, \dots, n\} \text{ is randomly chosen with uniform probability}$$
$$x_j^{k+1} = \text{prox}_{\gamma_j g_j}(x_j^k - \gamma_j \nabla f(x^k)_j)$$
$$x_i^{k+1} = x_i^k \text{ for all } i \neq j$$

in the convex setting for solving

$$\underset{x}{\text{minimize}} \quad f(x) + \underbrace{\sum_{i=1}^n g_i(x_i)}_{g(x)}$$

- Will show sublinear  $O(1/k)$  rate for expected function values
- Analysis based on *A fundamental inequality*

## Convex setting – Assumptions

(i)  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is continuously differentiable and convex

(ii)  $f$  is  $\beta_j$ -coordinate smooth, i.e., we have

$$f(y) \leq f(x) + \nabla f(x)_j (y_j - x_j) + \frac{\beta_j}{2} (x_j - y_j)^2$$

for all  $x, y \in \mathbb{R}^n$  such that  $y_i = x_i$  for all  $i \neq j$

(iii)  $g : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  closed convex and separable

(iv) A minimizer  $x^*$  exists and  $p^* = f(x^*) + g(x^*)$  is optimal value

(v) Algorithm parameters  $\gamma_j \in (0, \frac{1}{\beta_j}]$

- Same as for fundamental inequality but
  - restricted step-sizes
  - convexity of  $f$
- Smaller  $\gamma_j$  range than nonconvex, can be done with same range

## Convex setting – Analysis

- Use fundamental inequality with  $z = x^*$ , where  $x^*$  is a solution

$$\begin{aligned} & \mathbb{E}[f(x^{k+1}) + g(x^{k+1})|x^k] \\ & \leq f(x^k) + \frac{1}{n}g(x^*) + \frac{1}{n}\nabla f(x^k)^T(x^* - x^k) + \frac{n-1}{n}g(x^k) \\ & \quad + \frac{1}{2}\mathbb{E}[(\beta_j - \gamma_j^{-1})(x_j^{k+1} - x_j^k)^2|x^k] \\ & \quad + \frac{1}{2}(\mathbb{E}[\gamma_j^{-1}(x_j^k - x_j^*)^2|x^k] - \mathbb{E}[\gamma_j^{-1}(x_j^{k+1} - x_j^*)^2|x^k]) \end{aligned}$$

- Using  $\frac{1}{n}f(x^*) \geq \frac{1}{n}(f(x^k) + \nabla f(x^k)^T(x^* - x^k))$  by convexity of  $f$

$$\begin{aligned} & \mathbb{E}[f(x^{k+1}) + g(x^{k+1})|x^k] \\ & \leq \frac{n-1}{n}f(x^k) + \frac{1}{n}(g(x^*) + f(x^*)) + \frac{n-1}{n}g(x^k) \\ & \quad + \frac{1}{2}\mathbb{E}[(\beta_j - \gamma_j^{-1})(x_j^{k+1} - x_j^k)^2|x^k] \\ & \quad + \frac{1}{2}(\mathbb{E}[\gamma_j^{-1}(x_j^k - x_j^*)^2|x^k] - \mathbb{E}[\gamma_j^{-1}(x_j^{k+1} - x_j^*)^2|x^k]) \end{aligned}$$

## Analysis – Step-size requirement

- Restating what we just had

$$\begin{aligned} & \mathbb{E}[f(x^{k+1}) + g(x^{k+1})|x^k] \\ & \leq \frac{n-1}{n}f(x^k) + \frac{1}{n}(g(x^*) + f(x^*)) + \frac{n-1}{n}g(x^k) \\ & \quad + \frac{1}{2}\mathbb{E}[(\beta_j - \gamma_j^{-1})(x_j^{k+1} - x_j^k)^2|x^k] \\ & \quad + \frac{1}{2}(\mathbb{E}[\gamma_j^{-1}(x_j^k - x_j^*)^2|x^k] - \mathbb{E}[\gamma_j^{-1}(x_j^{k+1} - x_j^*)^2|x^k]) \end{aligned}$$

- Using  $\gamma_j \in (0, \frac{1}{\beta_j}]$  and  $p^* = f(x^*) + g(x^*)$ , rearrangement gives

$$\begin{aligned} & \frac{n-1}{n}\mathbb{E}[f(x^{k+1}) + g(x^{k+1})|x^k] + \frac{1}{2}\mathbb{E}[\gamma_j^{-1}(x_j^{k+1} - x_j^*)^2|x^k] \\ & \leq \frac{n-1}{n}(f(x^k) + g(x^k)) + \frac{1}{2}\mathbb{E}[\gamma_j^{-1}(x_j^k - x_j^*)^2|x^k] \\ & \quad - \frac{1}{n}(\mathbb{E}[f(x^{k+1}) + g(x^{k+1})|x^k] - p^*) \end{aligned}$$



## Lyapunov inequality

- Subtract  $\frac{n-1}{n}p^*$ , take expectation, use law of total expectation:

$$\begin{aligned} & \underbrace{\frac{n-1}{n}\mathbb{E}[f(x^{k+1}) + g(x^{k+1}) - p^*] + \frac{1}{2}\mathbb{E}[\gamma_j^{-1}(x_j^{k+1} - x_j^*)^2]}_{V_{k+1}} \\ & \leq \underbrace{\frac{n-1}{n}\mathbb{E}[f(x^k) + g(x^k) - p^*] + \frac{1}{2}\mathbb{E}[\gamma_j^{-1}(x_j^k - x_j^*)^2]}_{V_k} \\ & \quad - \underbrace{\frac{1}{n}(\mathbb{E}[f(x^{k+1}) + g(x^{k+1})] - p^*)}_{R_k} \end{aligned}$$

- Lyapunov inequality sequences  $V_k$  and  $R_k$  are nonnegative

## Lyapunov inequality consequences

- Lyapunov inequality  $V_{k+1} \leq V_k - R_k$  with

$$V_k = \frac{n-1}{n} \mathbb{E}[f(x^k) + g(x^k) - p^*] + \frac{1}{2} \mathbb{E}[\gamma_j^{-1} (x_j^k - x_j^*)^2]$$
$$R_k = \frac{1}{n} (\mathbb{E}[f(x^{k+1}) + g(x^{k+1})] - p^*)$$

and  $V_0 = \frac{n-1}{n} (f(x^0) + g(x^0) - p^*) + \frac{1}{2n} \|x^0 - x^*\|_{\Gamma^{-1}}^2$

- Consequences (similar to for proximal gradient method):
  - Since expected function value is decreasing:

$$\mathbb{E}[f(x^{k+1}) + g(x^{k+1})] - p^* \leq \frac{(n-1)(f(x^0) + g(x^0) - p^*) + \frac{1}{2} \|x^0 - x^*\|_{\Gamma^{-1}}^2}{k+1}$$

- Expected function value suboptimality summable

$$\sum_{l=0}^{\infty} \mathbb{E}[f(x^{l+1}) + g(x^{l+1}) - p^*] < \infty$$

so function value converges to  $p^*$  with probability 1

- Can show almost sure sequence convergence to an optimal point

# Outline

- Coordinate proximal gradient method
- Coordinate-wise smoothness
- A fundamental inequality
- Nonconvex setting
- Convex setting
- **Strongly convex setting**
- Rate comparison to proximal gradient method

## Strongly convex setting

- We will analyze the coordinate proximal gradient method

$$j \in \{1, \dots, n\} \text{ is randomly chosen with uniform probability}$$
$$x_j^{k+1} = \text{prox}_{\gamma_j g_j}(x_j^k - \gamma_j \nabla f(x^k)_j)$$
$$x_i^{k+1} = x_i^k \text{ for all } i \neq j$$

in a strongly convex setting for solving

$$\underset{x}{\text{minimize}} f(x) + \underbrace{\sum_{i=1}^n g_i(x_i)}_{g(x)}$$

- Will show linear convergence for  $\mathbb{E}[\|x^{k+1} - x^*\|_2]$
- Analysis based on properties of gradient

## Strongly convex setting – Assumptions

- (i)  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is continuously differentiable and  $\sigma$ -strongly convex
- (ii)  $f$  is  $\beta$  smooth
- (iii)  $g : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$  closed convex and separable
- (iv) A minimizer  $x^*$  exists and  $p^* = f(x^*) + g(x^*)$  is optimal value
- (v) Algorithm parameters  $\gamma_j = \gamma \in (0, \frac{2}{\beta})$

- Differs from assumption for fundamental inequality in
  - restricted step-sizes
  - strong convexity of  $f$
  - smoothness instead of coordinate-wise smoothness
- Will reduce analysis to analysis for proximal gradient method
- Analysis with coordinate-wise smoothness can improve rate

## Strongly convex setting – Analysis

Use that

- (a) the coordinate proximal gradient method, after selection of  $j$ , is:

$$x_j^{k+1} = (\text{prox}_{\gamma g}(x^k - \gamma \nabla f(x^k)))_j$$

- (b) the proximal gradient mapping satisfies in this setting

$$\|\text{prox}_{\gamma g}(x^k - \gamma \nabla f(x^k)) - x^*\|_2 \leq \max(1 - \sigma\beta, \beta\gamma - 1)\|x^k - x^*\|_2$$

to get

$$\begin{aligned}\mathbb{E}[\|x^{k+1} - x^*\|_2^2 | x^k] &= \mathbb{E}[(x_j^{k+1} - x_j^*)^2 | x^k] + \mathbb{E}\left[\sum_{i \neq j} (x_i^k - x_i^*)^2 | x^k\right] \\ &= \mathbb{E}[(\text{prox}_{\gamma g}(x^k - \gamma \nabla f(x^k)) - x^*)_j^2 | x^k] + \frac{n-1}{n} \|x^k - x^*\|_2^2 \\ &= \frac{1}{n} \|\text{prox}_{\gamma g}(x^k - \gamma \nabla f(x^k)) - x^*\|_2^2 + \frac{n-1}{n} \|x^k - x^*\|_2^2 \\ &\leq \frac{1}{n} \max(1 - \sigma\beta, \beta\gamma - 1)^2 \|x^k - x^*\|_2^2 + \frac{n-1}{n} \|x^k - x^*\|_2^2 \\ &\leq \left(1 - \frac{1}{n}(1 - \max(1 - \sigma\gamma, \beta\gamma - 1)^2)\right) \|x^k - x^*\|_2^2\end{aligned}$$

## Analysis – Total expectation

- Taking expectation and using law of total expectation gives

$$\mathbb{E}[\|x^{k+1} - x^*\|_2^2] \leq \underbrace{\left(1 - \frac{1}{n}(1 - \max(1 - \sigma\gamma, \beta\gamma - 1)^2)\right)}_{\rho} \mathbb{E}[\|x^k - x^*\|_2^2]$$

- Consequences:

- $\mathbb{E}[\|x^k - x^*\|_2^2]$  converges linearly whenever

$$\max(1 - \sigma\gamma, \beta\gamma - 1)^2 \in [0, 1)$$

which is same condition as for proximal gradient method

- Since expected value is summable,

$$\sum_{l=0}^k \mathbb{E}[\|x^l - x^*\|_2^2] \leq \frac{\|x^0 - x^*\|_2^2}{1 - \rho} < \infty$$

algorithm realizations converge to  $x^*$  with probability 1

# Outline

- Coordinate proximal gradient method
- Coordinate-wise smoothness
- A fundamental inequality
- Nonconvex setting
- Convex setting
- Strongly convex setting
- **Rate comparison to proximal gradient method**



## Comparison to proximal gradient method

Setting	Quantity	Proximal	Coordinate
Nonconvex	$\ \nabla f(\bar{x}^k)\ _2^2$	$O(1/k)$	$O(1/k)$
Convex	$f(x_k) + g(x_k) - p^*$	$O(1/k)$	$O(1/k)$
Strongly convex	$\ x_k - x^*\ _2$	$O(\rho_{\text{pg}}^k)$	$O(\rho_{\text{cpg}}^k)$

- Same order of magnitude in convergence for all classes
- Compare constants or linear rate to decide which is faster
- Will compare for convex and strongly convex settings assuming:
  - Problem dimension  $n$ :  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  and  $g : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$
  - That  $n$  coordinate steps at cost of 1 full step

## Comparison – Convex setting

- Assume  $nk$  coordinate steps at cost of  $k$  full steps
- Assume in the different setups:

(a)  $f$  is  $\beta_j$ -coordinate smooth and  $\gamma_j = \frac{1}{\beta_j}$

(b)  $f$  is  $\beta$ -smooth and  $\gamma = \frac{1}{\beta}$

(c)  $f$  is  $\beta_H$ -smooth w.r.t.  $\|\cdot\|_H$  and  $\gamma = \frac{1}{\beta_H}$

- Assume (a): Rate for  $nk$  coordinate proximal gradient steps

$$\mathbb{E}[f(x^{nk+1}) + g(x^{nk+1})] - p^* \leq \frac{(n-1)(f(x^0) + g(x^0) - p^*) + \frac{1}{2}\|x^0 - x^*\|_B^2}{nk+1}$$

where  $\Gamma = \mathbf{diag}(\gamma_1, \dots, \gamma_n)$  and  $B = \Gamma^{-1} = \mathbf{diag}(\beta_1, \dots, \beta_n)$

- Assume (b): Rate for  $k$  full proximal gradient steps

$$f(x^{k+1}) + g(x^{k+1}) - p^* \leq \frac{\beta\|x^0 - x^*\|_B^2}{2(k+1)}$$

- Assume (c): Rate for  $k$  full proximal gradient steps

$$f(x^{k+1}) + g(x^{k+1}) - p^* \leq \frac{\beta_H\|x^0 - x^*\|_H^2}{2(k+1)}$$

## Step-sizes for quadratics

- Consider convex  $f(x) = \frac{1}{2}x^T P x + q^T x$  and  $g = 0$
- Coordinate descent under Assumption (a)
  - Have shown  $\beta_j = p_{jj}$ -coordinate smoothness
  - So  $B = \mathbf{diag}(P)$  and coordinate update:

$$x_j^{k+1} = (\text{prox}_g^B(x^k - B^{-1}\nabla f(x^k)))_j$$

- Full proximal gradient under Assumption (b)
  - Have  $\beta = \lambda_{\max}(P)$ -smoothness
  - Algorithm

$$x^{k+1} = \text{prox}_{\frac{1}{\beta}g}(x^k - \frac{1}{\beta}\nabla f(x^k))$$

- Full scaled proximal gradient under Assumption (c)
  - Use same scaling as in coordinate case  $H = B = \mathbf{diag}(P)$
  - Algorithm

$$x^{k+1} = \text{prox}_{\frac{1}{\beta_B}g}^B(x^k - \frac{1}{\beta_B}B^{-1}\nabla f(x^k))$$

- Same step-length as coordinate if  $\beta_B = 1$

## Quantifying example – Step-sizes

- We generate  $P$  and  $q$  in  $f(x) = \frac{1}{2}x^T P x + q^T x$  as follows:
  - $P = C^T C$  and  $C \in \mathbb{R}^{20 \times 100}$  and all  $c_{ij} \sim \mathcal{N}(0, 1)$
  - $q_i \sim \mathcal{N}(0, 1)$

- Coordinate method and Assumption (a):  $\beta_j \in [10, 43]$
- Full method and Assumption (b):  $\beta = 193$
- Full method and Assumption (c): What is  $\beta_H = \beta_B$ ?
  - Since  $f$  quadratic with Hessian  $P$ , we have

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} \|x - y\|_P^2$$

- So  $f$  is  $\beta_B$ -smooth if  $\beta_B B = \beta_B \mathbf{diag}(P) \succeq P$ , since then:

$$f(y) - (f(x) + \nabla f(x)^T (y - x)) = \frac{1}{2} \|x - y\|_P^2 \leq \frac{\beta_B}{2} \|x - y\|_{\mathbf{diag}(P)}^2$$

which in this example holds for  $\beta_B = 9.1$

- Individual smoothness parameters satisfy  $\beta_B \beta_j \in [91, 392]$
- Step-sizes are inverse of  $\beta$ s, much longer steps in coordinate case

## Rates for quadratics

- Consider again convex  $f(x) = \frac{1}{2}x^T P x + q^T x$  and  $g = 0$
- Coordinate upper bound (with  $g = 0$ ) after  $nk$  iterations

$$\begin{aligned} \frac{(n-1)(f(x^0) - p^*) + \frac{1}{2}\|x^0 - x^*\|_B^2}{nk+1} &= \frac{\frac{(n-1)}{2}\|x^0 - x^*\|_P^2 + \frac{1}{2}\|x^0 - x^*\|_B^2}{nk+1} \\ &\approx \frac{n\|x^0 - x^*\|_P^2}{2(nk+1)} \approx \frac{\|x^0 - x^*\|_P^2}{2(k+1)} \end{aligned}$$

- Full and scaled proximal gradient upper bounds after  $k$  iterations:

$$\frac{\lambda_{\max}(P)\|x^0 - x^*\|_2^2}{2(k+1)} \qquad \frac{\beta_B\|x^0 - x^*\|_B^2}{2(k+1)}$$

- We know that rates are the same, but constants differ

## Quantifying example – Rate constants

- Quantify rate constants with same convex quadratic as before
- Coordinate, full, and scaled full proximal gradient rate constants:

$$\|x^0 - x^*\|_P^2 \quad \lambda_{\max}(P)\|x^0 - x^*\|_2^2 \quad \beta_B\|x^0 - x^*\|_B^2$$

- First two constants equal if  $x^0 - x^*$  is eigenvector to  $\lambda_{\max}(P)$
- Quantification: average constants ( $\overline{X}$ ) for  $N = 10000$  random  $x^0$

$$\overline{\|x^0 - x^*\|_P^2} \approx 2100$$

$$\overline{193\|x^0 - x^*\|_2^2} \approx 19300$$

$$\overline{9.1\|x^0 - x^*\|_{\mathbf{diag}(P)}^2} \approx 18900$$

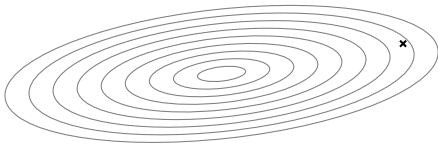
- Conclusions:
  - Coordinate does not improve worst case, but average performance
  - Coordinate descent almost 10 times smaller average constant here
  - No improvement in using  $\mathbf{diag}(P)$  for full method in this example

## Comparison – Toy example

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameters  $\gamma_1 = \frac{1}{0.1}$ ,  $\gamma_2 = 1$

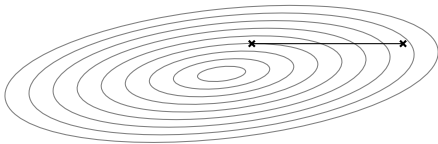


## Comparison – Toy example

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameters  $\gamma_1 = \frac{1}{0.1}$ ,  $\gamma_2 = 1$



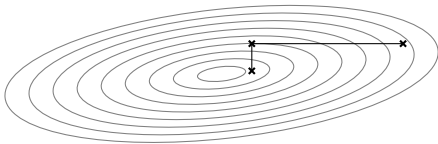


## Comparison – Toy example

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameters  $\gamma_1 = \frac{1}{0.1}$ ,  $\gamma_2 = 1$

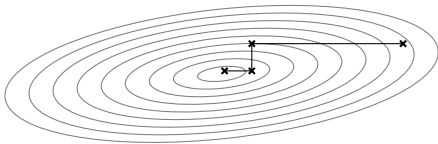


## Comparison – Toy example

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameters  $\gamma_1 = \frac{1}{0.1}$ ,  $\gamma_2 = 1$

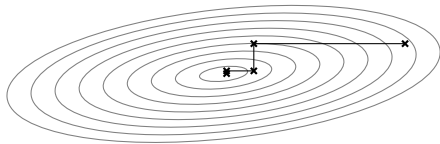


## Comparison – Toy example

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameters  $\gamma_1 = \frac{1}{0.1}$ ,  $\gamma_2 = 1$

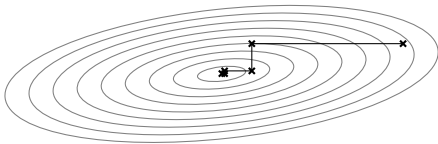


## Comparison – Toy example

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameters  $\gamma_1 = \frac{1}{0.1}$ ,  $\gamma_2 = 1$

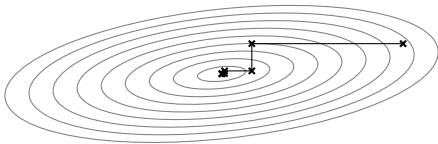


## Comparison – Toy example

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameters  $\gamma_1 = \frac{1}{0.1}$ ,  $\gamma_2 = 1$

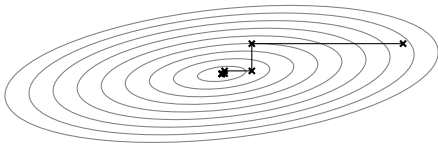


## Comparison – Toy example

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameters  $\gamma_1 = \frac{1}{0.1}$ ,  $\gamma_2 = 1$

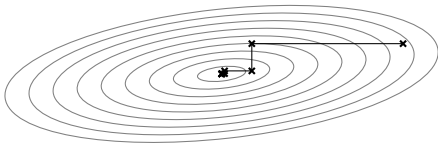


## Comparison – Toy example

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameters  $\gamma_1 = \frac{1}{0.1}$ ,  $\gamma_2 = 1$

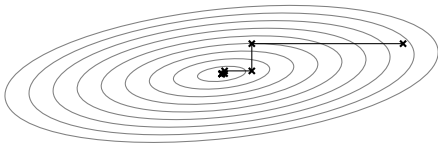


## Comparison – Toy example

- Coordinate descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameters  $\gamma_1 = \frac{1}{0.1}$ ,  $\gamma_2 = 1$



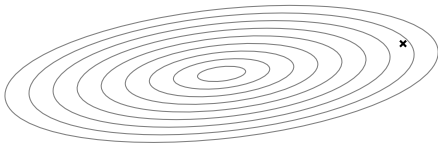


## Toy example – Gradient descent

- Gradient descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameter  $\gamma = \frac{1}{\beta}$

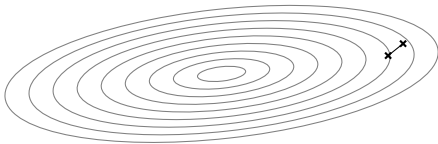


## Toy example – Gradient descent

- Gradient descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameter  $\gamma = \frac{1}{\beta}$

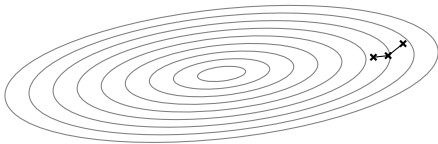


## Toy example – Gradient descent

- Gradient descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameter  $\gamma = \frac{1}{\beta}$

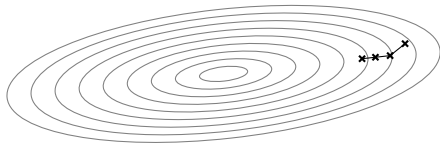


## Toy example – Gradient descent

- Gradient descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameter  $\gamma = \frac{1}{\beta}$

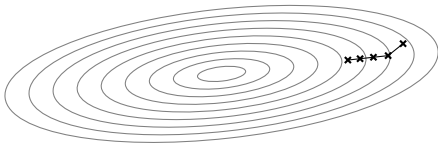


## Toy example – Gradient descent

- Gradient descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameter  $\gamma = \frac{1}{\beta}$

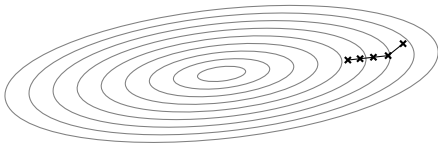


## Toy example – Gradient descent

- Gradient descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameter  $\gamma = \frac{1}{\beta}$

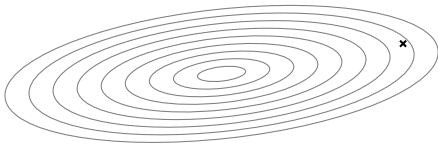


## Toy Example – Scaled gradient descent

- Diagonal scaled gradient descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameters  $\gamma_1 = \frac{1}{0.1\beta_H}$  and  $\gamma_2 = \frac{1}{\beta_H}$  with  $\beta_H = 1.32$

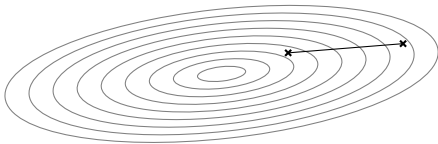


## Toy Example – Scaled gradient descent

- Diagonal scaled gradient descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameters  $\gamma_1 = \frac{1}{0.1\beta_H}$  and  $\gamma_2 = \frac{1}{\beta_H}$  with  $\beta_H = 1.32$



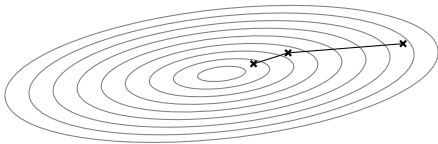


## Toy Example – Scaled gradient descent

- Diagonal scaled gradient descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameters  $\gamma_1 = \frac{1}{0.1\beta_H}$  and  $\gamma_2 = \frac{1}{\beta_H}$  with  $\beta_H = 1.32$

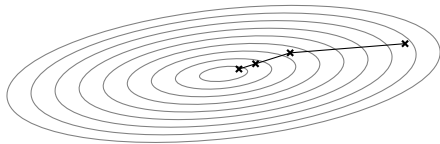


## Toy Example – Scaled gradient descent

- Diagonal scaled gradient descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameters  $\gamma_1 = \frac{1}{0.1\beta_H}$  and  $\gamma_2 = \frac{1}{\beta_H}$  with  $\beta_H = 1.32$

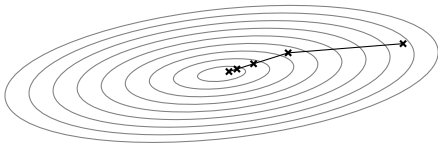


## Toy Example – Scaled gradient descent

- Diagonal scaled gradient descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameters  $\gamma_1 = \frac{1}{0.1\beta_H}$  and  $\gamma_2 = \frac{1}{\beta_H}$  with  $\beta_H = 1.32$

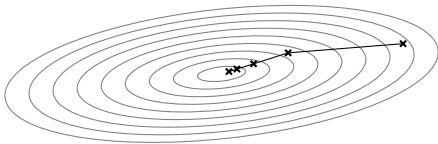


## Toy Example – Scaled gradient descent

- Diagonal scaled gradient descent on  $\beta$ -smooth quadratic problem

$$\underset{x}{\text{minimize}} \quad \frac{1}{2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^T \begin{bmatrix} 0.1 & -0.1 \\ -0.1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Step-size parameters  $\gamma_1 = \frac{1}{0.1\beta_H}$  and  $\gamma_2 = \frac{1}{\beta_H}$  with  $\beta_H = 1.32$



## Comparison – Strongly convex setting

- Assumptions:
  - $nk$  coordinate steps at cost of  $k$  full steps
  - All step-sizes fixed to be the same, also in coordinate
- Rates for  $k$  proximal and  $nk$  coordinate proximal steps

$$\|x_k - x^*\|_2 \leq \max(\beta\gamma - 1, 1 - \sigma\gamma)^k \|x_0 - x^*\|_2$$

$$\mathbb{E}[\|x_{kn} - x^*\|_2] \leq \left(1 - \frac{1}{n}(1 - \max(\beta\gamma - 1, 1 - \sigma\gamma))^2\right)^{nk/2} \|x_0 - x^*\|_2$$

## Strongly convex comparison – Example

- Comparison on  $f(x) = \frac{1}{2}x^T Px + q^T x$  and arbitrary convex  $g$ 
  - $P = C^T C$  and  $C \in \mathbb{R}^{100 \times 100}$  and all  $c_{ij} \in \mathcal{N}(0, 1)$
  - We have  $\beta = \lambda_{\max}(P) \approx 399$  and  $\sigma = \lambda_{\min}(P) \approx 0.007$
  - We let  $\gamma = \frac{1}{\beta}$  and compare for  $k = 10000$  steps (epocs)

$$(\beta\gamma - 1, 1 - \sigma\gamma)^k \approx 0.837686$$

$$(1 - \frac{1}{n}(1 - \max(\beta\gamma - 1, 1 - \sigma\gamma)))^2)^{nk/2} \approx 0.837689$$

- Comments:
  - With identical step-sizes, rates are very similar
  - Coordinate method can take longer steps to get better rate (but not covered by our strongly convex analysis)