CS7301: Advanced Topics in Optimization for Machine Learning

Lecture 5.1: Second Order Methods Cont. and Coordinate Descent

Rishabh Iyer

Department of Computer Science
University of Texas, Dallas
https://github.com/rishabhk108/AdvancedOptML

February 26, 2021

Summary of Algorithms so Far

First Order Methods (Unconstrained)

- Gradient Descent
- Proximal Gradient Descent (if non-smooth function is simple – i.e. proximal operator of that function can be computed efficiently.
- Accelerated Gradient versions in the unconstrained case, projected case and proximal case.

First Order Methods (Constrained)

- Projected Gradient Descent (whenever the Projection step is easy)
- Conditional Gradient Descent

Last Class

Second Order Methods

- Newtons Method
- General Quasi Newton Template
- Symmetric Rank One Update
- BFGS Update
- DFP Update

Outline of Lecture 5.1

- LBFGS (Limited Memory BFGS)
- Conjugate Gradient Descent
- Barzelia Borowein Gradient Descent
- Coordinate Descent Algorithms

Recall: BFGS Update

- BFGS update is a rank two update:

$$B_k = B_{k-1} + auu^T + bvv^T$$

- Recall BFGS update:

$$B_k = B_{k-1} - \frac{B_{k-1}s_{k-1}s_{k-1}^T B_{k-1}}{s_{k-1}^T B_{k-1}s_{k-1}} + \frac{y_{k-1}y_{k-1}^T}{y_{k-1}^T s_{k-1}}$$

- BFGS Inverse Update:

$$B_{k+1}^{-1} = (I - \frac{s_{k-1}y_{k-1}^T}{s_{k-1}^T y_{k-1}})B_k^{-1}(I - \frac{y_{k-1}s_{k-1}^T}{y_{k-1}^T s_{k-1}}) + \frac{s_{k-1}s_{k-1}^T}{y_{k-1}^T s_{k-1}}$$

Limited Memory BFGS

- The memory and run-time cost of $O(d^2)$ is itself not scalable for large number of features!

Limited Memory BFGS

- The memory and run-time cost of $O(d^2)$ is itself not scalable for large number of features!
- We need the improved (e.g. super-linear) convergence of BFGS/Newton but with linear or sublinear per iteration complexity!

Limited Memory BFGS

- The memory and run-time cost of $O(d^2)$ is itself not scalable for large number of features!
- We need the improved (e.g. super-linear) convergence of BFGS/Newton but with linear or sublinear per iteration complexity!
- Idea of LBFGS: Do not store B_k^{-1} explicitly.

Limited Memory BFGS

- The memory and run-time cost of $O(d^2)$ is itself not scalable for large number of features!
- We need the improved (e.g. super-linear) convergence of BFGS/Newton but with linear or sublinear per iteration complexity!
- Idea of LBFGS: Do not store B_k^{-1} explicitly.
- Instead store upto l (e.g. $l = 30$) values of s_k and y_k and compute B_k^{-1} using the recursive BFGS equation and assuming $B_{k-l} = I$:

$$B_{k+1}^{-1} = (I - \frac{s_{k-1}y_{k-1}^T}{s_{k-1}^T y_{k-1}})B_k^{-1}(I - \frac{y_{k-1}s_{k-1}^T}{y_{k-1}^T s_{k-1}}) + \frac{s_{k-1}s_{k-1}^T}{y_{k-1}^T s_{k-1}}$$

Limited Memory BFGS

- The memory and run-time cost of $O(d^2)$ is itself not scalable for large number of features!
- We need the improved (e.g. super-linear) convergence of BFGS/Newton but with linear or sublinear per iteration complexity!
- Idea of LBFGS: Do not store B_k^{-1} explicitly.
- Instead store upto l (e.g. $l = 30$) values of s_k and y_k and compute B_k^{-1} using the recursive BFGS equation and assuming $B_{k-l} = I$:

$$B_{k+1}^{-1} = (I - \frac{s_{k-1}y_{k-1}^T}{s_{k-1}^T y_{k-1}})B_k^{-1}(I - \frac{y_{k-1}s_{k-1}^T}{y_{k-1}^T s_{k-1}}) + \frac{s_{k-1}s_{k-1}^T}{y_{k-1}^T s_{k-1}}$$

- This makes the cost per iteration of $O(dl)$ and storage also as $O(dl)$ instead of $O(d^2)$.

Summary so Far...

- Newton's method: Faster than any other algorithm seen so far. However costly $O(d^3)$ to Invert Hessian

Summary so Far...

- Newton's method: Faster than any other algorithm seen so far. However costly $O(d^3)$ to Invert Hessian
- SR1, BFGS and DFP are simpler updates (uwing the Secant condition) which are Newton like (hence Quasi-Newton) but do this in $O(d^2)$ complexity!

Summary so Far...

- Newton's method: Faster than any other algorithm seen so far. However costly $O(d^3)$ to Invert Hessian
- SR1, BFGS and DFP are simpler updates (uwing the Secant condition) which are Newton like (hence Quasi-Newton) but do this in $O(d^2)$ complexity!
- LBFGS is an approximation to BFGS to make the algorithm linear!

Summary so Far...

- Newton's method: Faster than any other algorithm seen so far. However costly $O(d^3)$ to Invert Hessian
- SR1, BFGS and DFP are simpler updates (uwing the Secant condition) which are Newton like (hence Quasi-Newton) but do this in $O(d^2)$ complexity!
- LBFGS is an approximation to BFGS to make the algorithm linear!
- Next: Can we achieve Newton like behaviour with Gradient Descent like algorithms?

Conjugate Gradient Descent

- Linear Conjugate Gradient is an important Algorithm for quadratic optimization

Conjugate Gradient Descent

- Linear Conjugate Gradient is an important Algorithm for quadratic optimization
- Conjugate gradient solves the problem $Ax = b$ (which can be seen as minimizing $\frac{1}{2}x^T Ax - b^T x$)

Conjugate Gradient Descent

- Linear Conjugate Gradient is an important Algorithm for quadratic optimization
- Conjugate gradient solves the problem $Ax = b$ (which can be seen as minimizing $\frac{1}{2}x^T Ax - b^T x$)
- Non-Linear Conjugate Gradient is an extension to general convex functions!

Conjugate Gradient Descent

- Linear Conjugate Gradient is an important Algorithm for quadratic optimization
- Conjugate gradient solves the problem $Ax = b$ (which can be seen as minimizing $\frac{1}{2}x^T Ax - b^T x$)
- Non-Linear Conjugate Gradient is an extension to general convex functions!
- Several variants of conjugate gradient descent. Basic idea: Instead of using $-\nabla f(x_k)$ as the update direction (as in gradient descent), do the following:

Conjugate Gradient Descent

- Linear Conjugate Gradient is an important Algorithm for quadratic optimization
- Conjugate gradient solves the problem $Ax = b$ (which can be seen as minimizing $\frac{1}{2}x^T Ax - b^T x$)
- Non-Linear Conjugate Gradient is an extension to general convex functions!
- Several variants of conjugate gradient descent. Basic idea: Instead of using $-\nabla f(x_k)$ as the update direction (as in gradient descent), do the following:
  - Set $x_{k+1} = x_k + \alpha_k d_k$

Conjugate Gradient Descent

- Linear Conjugate Gradient is an important Algorithm for quadratic optimization
- Conjugate gradient solves the problem $Ax = b$ (which can be seen as minimizing $\frac{1}{2}x^T Ax - b^T x$)
- Non-Linear Conjugate Gradient is an extension to general convex functions!
- Several variants of conjugate gradient descent. Basic idea: Instead of using $-\nabla f(x_k)$ as the update direction (as in gradient descent), do the following:
  - Set $x_{k+1} = x_k + \alpha_k d_k$
  - Set $d_{k+1} = -g_{k+1} + \beta_k d_k$ where β_k is appropriately set!

Conjugate Gradient Descent for Quadratic Functions

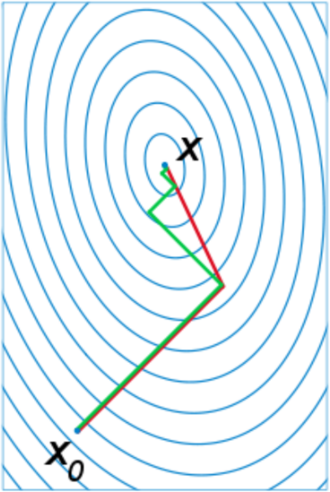
A comparison of

* gradient descent with optimal step size (in green) and

* conjugate vector (in red)

for minimizing a quadratic function.

Conjugate gradient converges in at most n steps (here n=2).



Variants of Conjugate gradient Descent

- Conjugate Gradient Framework:

Variants of Conjugate gradient Descent

- Conjugate Gradient Framework:
  - Set $x_{k+1} = x_k + \alpha_k d_k$

Variants of Conjugate gradient Descent

- Conjugate Gradient Framework:
  - Set $x_{k+1} = x_k + \alpha_k d_k$
  - Set $d_{k+1} = -g_{k+1} + \beta_k d_k$ where β_k is appropriately set!

Variants of Conjugate gradient Descent

- Conjugate Gradient Framework:
  - Set $x_{k+1} = x_k + \alpha_k d_k$
  - Set $d_{k+1} = -g_{k+1} + \beta_k d_k$ where β_k is appropriately set!
- Fletcher–Reeves CG algorithm:

$$\beta_k = \frac{||\nabla f(x_k)||^2}{||\nabla f(x_{k-1})||^2}$$

Variants of Conjugate gradient Descent

- Conjugate Gradient Framework:
  - Set $x_{k+1} = x_k + \alpha_k d_k$
  - Set $d_{k+1} = -g_{k+1} + \beta_k d_k$ where β_k is appropriately set!
- Fletcher–Reeves CG algorithm:

$$\beta_k = \frac{||\nabla f(x_k)||^2}{||\nabla f(x_{k-1})||^2}$$

- Polak–Ribière:

$$\beta_k = \frac{\nabla f(x_k)^T(\nabla f(x_k) - \nabla f(x_{k-1}))}{||\nabla f(x_{k-1})||^2}$$

Variants of Conjugate gradient Descent

- Conjugate Gradient Framework:
  - Set $x_{k+1} = x_k + \alpha_k d_k$
  - Set $d_{k+1} = -g_{k+1} + \beta_k d_k$ where β_k is appropriately set!
- Fletcher–Reeves CG algorithm:

$$\beta_k = \frac{||\nabla f(x_k)||^2}{||\nabla f(x_{k-1})||^2}$$

- Polak–Ribière:

$$\beta_k = \frac{\nabla f(x_k)^T(\nabla f(x_k) - \nabla f(x_{k-1}))}{||\nabla f(x_{k-1})||^2}$$

- Hestenes–Stiefel:

$$\beta_k = \frac{\nabla f(x_k)^T(\nabla f(x_k) - \nabla f(x_{k-1})}{d_k^T(\nabla f(x_k) - \nabla f(x_{k-1})}$$

- Hestenes–Stiefel conjugate gradient update is equivalent to LBFGS with $l = 1$!

Conjugate Gradient and Quasi Newton

- Hestenes–Stiefel conjugate gradient update is equivalent to LBFGS with $l = 1$!
- Moreover, similar to Newton and BFGS, one can also show superlinear local convergence and global convergence results for CG!

Conjugate Gradient and Quasi Newton

- Hestenes–Stiefel conjugate gradient update is equivalent to LBFGS with $l = 1$!
- Moreover, similar to Newton and BFGS, one can also show superlinear local convergence and global convergence results for CG!
- Advantage of CG is that the complexity is exactly identical to Gradient Descent and yet it has some of the properties of Quasi Newton algorithms!

Hessian Free Optimization (James Martin, ICML 2010)

- Recall the quadratic expansion:

$$f(x) \approx f(x_t) + \nabla f(x_t)^T (x - x_t) + \frac{1}{2}(x - x_t)^T \nabla^2 f(x_t)(x - x_t)$$

Hessian Free Optimization (James Martin, ICML 2010)

- Recall the quadratic expansion:

$$f(x) \approx f(x_t) + \nabla f(x_t)^T(x - x_t) + \frac{1}{2}(x - x_t)^T \nabla^2 f(x_t)(x - x_t)$$

- Compute x_{t+1} by optimizing the quadratic function above using the (linear) conjugate gradient!

Hessian Free Optimization (James Martin, ICML 2010)

- Recall the quadratic expansion:

$$f(x) \approx f(x_t) + \nabla f(x_t)^T (x - x_t) + \frac{1}{2}(x - x_t)^T \nabla^2 f(x_t)(x - x_t)$$

- Compute x_{t+1} by optimizing the quadratic function above using the (linear) conjugate gradient!
- Repeat these steps till converged

Hessian Free Optimization (James Martin, ICML 2010)

- Recall the quadratic expansion:

$$f(x) \approx f(x_t) + \nabla f(x_t)^T(x - x_t) + \frac{1}{2}(x - x_t)^T \nabla^2 f(x_t)(x - x_t)$$

- Compute x_{t+1} by optimizing the quadratic function above using the (linear) conjugate gradient!
- Repeat these steps till converged
- Moreover, we can do this Hessian Free(!) because we just need to compute $\nabla^2 f(x_t)v$, i.e. we don't need to compute $\nabla^2 f$ for this!

Hessian Free Optimization (James Martin, ICML 2010)

- Recall the quadratic expansion:

$$f(x) \approx f(x_t) + \nabla f(x_t)^T (x - x_t) + \frac{1}{2}(x - x_t)^T \nabla^2 f(x_t)(x - x_t)$$

- Compute x_{t+1} by optimizing the quadratic function above using the (linear) conjugate gradient!
- Repeat these steps till converged
- Moreover, we can do this Hessian Free(!) because we just need to compute $\nabla^2 f(x_t)v$, i.e. we don't need to compute $\nabla^2 f$ for this!
- Note that $(\nabla^2 f(x)v)_i = \sum_{j=1}^d \frac{\partial^2 f}{\partial x_i \partial x_j}(x).v_j = \nabla \frac{\partial f}{\partial x_i}(x).v$

Hessian Free Optimization (James Martin, ICML 2010)

- Recall the quadratic expansion:

$$f(x) \approx f(x_t) + \nabla f(x_t)^T (x - x_t) + \frac{1}{2}(x - x_t)^T \nabla^2 f(x_t)(x - x_t)$$

- Compute x_{t+1} by optimizing the quadratic function above using the (linear) conjugate gradient!
- Repeat these steps till converged
- Moreover, we can do this Hessian Free(!) because we just need to compute $\nabla^2 f(x_t)v$, i.e. we don't need to compute $\nabla^2 f$ for this!
- Note that $(\nabla^2 f(x)v)_i = \sum_{j=1}^d \frac{\partial^2 f}{\partial x_i \partial x_j}(x).v_j = \nabla \frac{\partial f}{\partial x_i}(x).v$
- This is the directional derivative of $\frac{\partial f}{\partial x_i}$ in the direction of v! and can be computed numerically!

Barzelai Borwein (BB) Gradient Descent

- The BB Method was introduced in an 8-page paper in 1988 (J. Barzilao and J. Borwein. Two point step size gradient method, IMA J. Numerical Analysis, 1988)

- The BB Method was introduced in an 8-page paper in 1988 (J. Barzilao and J. Borwein. Two point step size gradient method, IMA J. Numerical Analysis, 1988)
- Its a Gradient method with special step sizes.

Barzelai Borwein (BB) Gradient Descent

- The BB Method was introduced in an 8-page paper in 1988 (J. Barzilao and J. Borwein. Two point step size gradient method, IMA J. Numerical Analysis, 1988)
- Its a Gradient method with special step sizes.
- Motivated by Newtons method but does not compute the Hessian!

Barzelai Borwein (BB) Gradient Descent

- The BB Method was introduced in an 8-page paper in 1988 (J. Barzilao and J. Borwein. Two point step size gradient method, IMA J. Numerical Analysis, 1988)
- Its a Gradient method with special step sizes.
- Motivated by Newtons method but does not compute the Hessian!
- At no extra cost over the standard gradient descent, this method can significantly outperform GD!

Barzelai Borwein (BB) Gradient Descent

- The BB Method was introduced in an 8-page paper in 1988 (J. Barzilao and J. Borwein. Two point step size gradient method, IMA J. Numerical Analysis, 1988)
- Its a Gradient method with special step sizes.
- Motivated by Newtons method but does not compute the Hessian!
- At no extra cost over the standard gradient descent, this method can significantly outperform GD!
- Typically used with non-monotone line search as convergence safeguard against non-quadratic problems (we won't cover this in class).

Barzelai Borwein (BB) Gradient Descent

- Recall Gradient Descent:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

. Pros: Simple! Cons: No second order information at all!

Barzelai Borwein (BB) Gradient Descent

- Recall Gradient Descent:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

. Pros: Simple! Cons: No second order information at all!

- Newton's method (similarly BFGS):

$$x_{t+1} = x_t - \alpha_t H_t^{-1} \nabla f(x_t)$$

. Pros: Very Fast Convergence! Cons: Creating or updating H_k^{-1} superlinear in cost!

Barzelai Borwein (BB) Gradient Descent

- Recall Gradient Descent:

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

. Pros: Simple! Cons: No second order information at all!

- Newton's method (similarly BFGS):

$$x_{t+1} = x_t - \alpha_t H_t^{-1} \nabla f(x_t)$$

. Pros: Very Fast Convergence! Cons: Creating or updating H_k^{-1} superlinear in cost!

- BB Gradient Descent: Choose $\alpha_t \nabla f(x_t)$ such that it approximates $H_t^{-1} \nabla f(x_t)$

Barzelai Borwein (BB) Gradient Descent

- Recall Secant Condition:

$$H_k s_{k-1} = y_{k-1}$$

where $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = \nabla f(x_k) - \nabla f(x_{k-1}$

Barzelai Borwein (BB) Gradient Descent

- Recall Secant Condition:

$$H_k s_{k-1} = y_{k-1}$$

where $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = \nabla f(x_k) - \nabla f(x_{k-1}$

- Another way to see the secant condition is the fact that the this is the expression of the Hessian for quadratic functions!

Barzelai Borwein (BB) Gradient Descent

- Recall Secant Condition:

$$H_k s_{k-1} = y_{k-1}$$

where $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = \nabla f(x_k) - \nabla f(x_{k-1}$

- Another way to see the secant condition is the fact that the this is the expression of the Hessian for quadratic functions!
- Now, the BB update is very simple: Select α_k such that:

$$(\alpha_k^{-1} I) s_{k-1} \approx y_{k-1}$$

Barzelai Borwein (BB) Gradient Descent

- Recall Secant Condition:

$$H_k s_{k-1} = y_{k-1}$$

where $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = \nabla f(x_k) - \nabla f(x_{k-1}$

- Another way to see the secant condition is the fact that the this is the expression of the Hessian for quadratic functions!
- Now, the BB update is very simple: Select α_k such that:

$$(\alpha_k^{-1} I) s_{k-1} \approx y_{k-1}$$

- Solve this using least squares!

- Goal: Select α_k such that:

$$(\alpha_k^{-1} I) s_{k-1} \approx y_{k-1}$$

Barzelai Borwein (BB) Gradient Descent

- Goal: Select α_k such that:

$$(\alpha_k^{-1} I) s_{k-1} \approx y_{k-1}$$

- BB Method:

Barzelai Borwein (BB) Gradient Descent

- Goal: Select α_k such that:

$$(\alpha_k^{-1} I) s_{k-1} \approx y_{k-1}$$

- BB Method:
  - Approach 1: Let $\beta = \alpha^{-1}$. Then:

$$\alpha_k^{-1} = \text{argmin}_\alpha \frac{1}{2} ||s_{k-1}\beta - y_{k-1}||^2 \Rightarrow \alpha_k = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}}$$

Barzelai Borwein (BB) Gradient Descent

- Goal: Select α_k such that:

$$(\alpha_k^{-1} I) s_{k-1} \approx y_{k-1}$$

- BB Method:
  - Approach 1: Let $\beta = \alpha^{-1}$. Then:

$$\alpha_k^{-1} = \text{argmin}_\alpha \frac{1}{2} ||s_{k-1}\beta - y_{k-1}||^2 \Rightarrow \alpha_k = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}}$$

  - Approach 2: Least squares in α directly:

$$\alpha_k = \text{argmin}_\alpha \frac{1}{2} ||s_{k-1} - y_{k-1}\alpha||^2 \Rightarrow \alpha_k = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}}$$

BB Gradient Descent in Practice

- BB Method:

BB Gradient Descent in Practice

- BB Method:
  - Approach 1: Let $\beta = \alpha^{-1}$. Then:

$$\alpha_k^{-1} = \text{argmin}_\alpha \frac{1}{2}||s_{k-1}\beta - y_{k-1}||^2 \Rightarrow \alpha_k = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}}$$

BB Gradient Descent in Practice

- BB Method:
  - Approach 1: Let $\beta = \alpha^{-1}$. Then:

$$\alpha_k^{-1} = \text{argmin}_\alpha \frac{1}{2}||s_{k-1}\beta - y_{k-1}||^2 \Rightarrow \alpha_k = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}}$$

  - Approach 2: Least squares in α directly:

$$\alpha_k = \text{argmin}_\alpha \frac{1}{2}||s_{k-1} - y_{k-1}\alpha||^2 \Rightarrow \alpha_k = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}}$$

BB Gradient Descent in Practice

- BB Method:
  - Approach 1: Let $\beta = \alpha^{-1}$. Then:

$$\alpha_k^{-1} = \text{argmin}_\alpha \frac{1}{2}||s_{k-1}\beta - y_{k-1}||^2 \Rightarrow \alpha_k = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}}$$

  - Approach 2: Least squares in α directly:

$$\alpha_k = \text{argmin}_\alpha \frac{1}{2}||s_{k-1} - y_{k-1}\alpha||^2 \Rightarrow \alpha_k = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}}$$

- In practice, use either of these two or pick the better among the two or even alternate by fixing one for a few steps.

BB Gradient Descent in Practice

- BB Method:
  - Approach 1: Let $\beta = \alpha^{-1}$. Then:

$$\alpha_k^{-1} = \text{argmin}_\alpha \frac{1}{2}||s_{k-1}\beta - y_{k-1}||^2 \Rightarrow \alpha_k = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}}$$

  - Approach 2: Least squares in α directly:

$$\alpha_k = \text{argmin}_\alpha \frac{1}{2}||s_{k-1} - y_{k-1}\alpha||^2 \Rightarrow \alpha_k = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}}$$

- In practice, use either of these two or pick the better among the two or even alternate by fixing one for a few steps.
- However, $f(x_k)$ and $\nabla f(x_k)$ are not monotonic!

BB Gradient Descent in Practice

- BB Method:
  - Approach 1: Let $\beta = \alpha^{-1}$. Then:

$$\alpha_k^{-1} = \text{argmin}_\alpha \frac{1}{2}||s_{k-1}\beta - y_{k-1}||^2 \Rightarrow \alpha_k = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}}$$

  - Approach 2: Least squares in α directly:

$$\alpha_k = \text{argmin}_\alpha \frac{1}{2}||s_{k-1} - y_{k-1}\alpha||^2 \Rightarrow \alpha_k = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}}$$

- In practice, use either of these two or pick the better among the two or even alternate by fixing one for a few steps.
- However, $f(x_k)$ and $\nabla f(x_k)$ are not monotonic!
- Non Monotonic line search required for convergence analysis.

Summary So Far...

- Gradient Descent is easy to implement, simple with low per iteration complexity. However, at best linear convergence! For most standard cases, sublinear convergence.

Summary So Far...

- Gradient Descent is easy to implement, simple with low per iteration complexity. However, at best linear convergence! For most standard cases, sublinear convergence.
- Accelerated Gradient Descent is an improvement. Slightly more involved but similar per iteration complexity. Slightly faster but still similar to GD!

Summary So Far...

- Gradient Descent is easy to implement, simple with low per iteration complexity. However, at best linear convergence! For most standard cases, sublinear convergence.
- Accelerated Gradient Descent is an improvement. Slightly more involved but similar per iteration complexity. Slightly faster but still similar to GD!
- Newton Method is superlinear convergence! However very slow per iterations: $O(d^3)$ complexity.

Summary So Far...

- Gradient Descent is easy to implement, simple with low per iteration complexity. However, at best linear convergence! For most standard cases, sublinear convergence.
- Accelerated Gradient Descent is an improvement. Slightly more involved but similar per iteration complexity. Slightly faster but still similar to GD!
- Newton Method is superlinear convergence! However very slow per iterations: $O(d^3)$ complexity.
- BFGS improves it to $O(d^2)$ which is further improved to $O(d)$ by LBFGS!

Summary So Far...

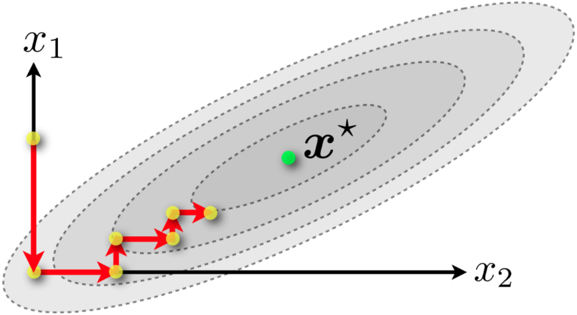
- Gradient Descent is easy to implement, simple with low per iteration complexity. However, at best linear convergence! For most standard cases, sublinear convergence.
- Accelerated Gradient Descent is an improvement. Slightly more involved but similar per iteration complexity. Slightly faster but still similar to GD!
- Newton Method is superlinear convergence! However very slow per iterations: $O(d^3)$ complexity.
- BFGS improves it to $O(d^2)$ which is further improved to $O(d)$ by LBFGS!
- CG and BB are variants of Gradient Descent but try to mimic Newton's methods and in practice are much faster than simple GD while maintaining the simplicity!

Coordinate Descent

-
- Coordinate Descent: Basic Idea
- Optimality w.r.t co-ordinate descent
- Two variants of Coordinate Descent
- Convergence of coordinate descent algorithms.

Coordinate Descent

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



Idea: Update one coordinate at a time, while keeping others fixed.

Coordinate Descent

- Modify only one coordinate per step:

Coordinate Descent

- Modify only one coordinate per step:
  1. Select $i_k \in [d]$

Coordinate Descent

- Modify only one coordinate per step:
  1. Select $i_k \in [d]$
  2. $x_{k+1} = x_k + \gamma e_{i_k}$

Coordinate Descent

- Modify only one coordinate per step:
  1. Select $i_k \in [d]$
  2. $x_{k+1} = x_k + \gamma e_{i_k}$
- γ is the step size of coordinate descent

Coordinate Descent

- Modify only one coordinate per step:
  1. Select $i_k \in [d]$
  2. $x_{k+1} = x_k + \gamma e_{i_k}$
- γ is the step size of coordinate descent
- Two variants of coordinate descent:

Coordinate Descent

- Modify only one coordinate per step:
  1. Select $i_k \in [d]$
  2. $x_{k+1} = x_k + \gamma e_{i_k}$
- γ is the step size of coordinate descent
- Two variants of coordinate descent:
- Gradient based step size:

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

Coordinate Descent

- Modify only one coordinate per step:
  1. Select $i_k \in [d]$
  2. $x_{k+1} = x_k + \gamma e_{i_k}$
- γ is the step size of coordinate descent
- Two variants of coordinate descent:
- Gradient based step size:

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

- Exact coordinate minimization: solve the single variable minimization

$$\text{argmin}_{\gamma \in R} f(x_k + \gamma e_{i_k})$$

in closed form.

Coordinate Descent

- Two variants of coordinate descent:

Coordinate Descent

- Two variants of coordinate descent:
- Gradient based step size:

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

Coordinate Descent

- Two variants of coordinate descent:
- Gradient based step size:

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

- Exact coordinate minimization: solve the single variable minimization

$$\text{argmin}_{\gamma \in R} f(x_k + \gamma e_{i_k})$$

in closed form.

Coordinate Descent

- Two variants of coordinate descent:
- Gradient based step size:

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

- Exact coordinate minimization: solve the single variable minimization

$$\text{argmin}_{\gamma \in R} f(x_k + \gamma e_{i_k})$$

  in closed form.
- Also how do we select the coordinate i_k?

Coordinate Descent

- Two variants of coordinate descent:
- Gradient based step size:

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

- Exact coordinate minimization: solve the single variable minimization

$$\text{argmin}_{\gamma \in R} f(x_k + \gamma e_{i_k})$$

  in closed form.
- Also how do we select the coordinate i_k?
- Two strategies: One is to pick it up randomly (called randomized coordinate descent) and second it to pick it up greedily (greedy coordinate descent).

Randomized Coordinate Descent

select $i_t \in [d]$ uniformly at random
$$\mathbf{x}_{t+1} := \mathbf{x}_t - \frac{1}{L}\nabla_{i_t} f(\mathbf{x}_t)\, \mathbf{e}_{i_t}$$

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

Randomized Coordinate Descent: Convergence Analysis

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

Randomized Coordinate Descent: Convergence Analysis

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- This is equivalent to coordinate wise Lipschitz gradient:

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

Randomized Coordinate Descent: Convergence Analysis

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- This is equivalent to coordinate wise Lipschitz gradient:

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

- Define $L = \max_i L_i$ as the Maximum coordinate wise Lipschitz smoothness

Randomized Coordinate Descent: Convergence Analysis

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- This is equivalent to coordinate wise Lipschitz gradient:

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

- Define $L = \max_i L_i$ as the Maximum coordinate wise Lipschitz smoothness
- Similar to previous analysis, lets assume $||x_0 - x^*||^2 \leq R^2$.

Randomized Coordinate Descent: Convergence Result

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

Randomized Coordinate Descent: Convergence Result

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- Define $L = \max_i L_i$ as the Maximum coordinate wise Lipschitz smoothness. Also, assume $||x_0 - x^*||^2 \leq R^2$.

Randomized Coordinate Descent: Convergence Result

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- Define $L = \max_i L_i$ as the Maximum coordinate wise Lipschitz smoothness. Also, assume $||x_0 - x^*||^2 \leq R^2$.
- **Theorem:** For a coordinate wise smooth convex function, we have:

$$E(f(x_k)) - f_* \leq \frac{2dLR_0^2}{k}$$

Randomized Coordinate Descent: Convergence Result

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- Define $L = \max_i L_i$ as the Maximum coordinate wise Lipschitz smoothness. Also, assume $||x_0 - x^*||^2 \leq R^2$.
- **Theorem:** For a coordinate wise smooth convex function, we have:

$$E(f(x_k)) - f_* \leq \frac{2dLR_0^2}{k}$$

- Similar to Gradient Descent, this is $O(1/\epsilon)$ convergence!

Randomized Coordinate Descent: Convergence Result

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- Define $L = \max_i L_i$ as the Maximum coordinate wise Lipschitz smoothness. Also, assume $||x_0 - x^*||^2 \leq R^2$.
- **Theorem:** For a coordinate wise smooth convex function, we have:

$$E(f(x_k)) - f_* \leq \frac{2dLR_0^2}{k}$$

- Similar to Gradient Descent, this is $O(1/\epsilon)$ convergence!
- However, we just need to update a single dimension at a time, and hence the per iteration cost of CD can be $O(d)$ cheaper compared to GD!

Randomized Coordinate Descent: Convergence Result

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

Randomized Coordinate Descent: Convergence Result

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- Define $L = \max_i L_i$ as the Maximum coordinate wise Lipschitz smoothness. Also, assume $||x_0 - x^*||^2 \leq R^2$.

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- Define $L = \max_i L_i$ as the Maximum coordinate wise Lipschitz smoothness. Also, assume $||x_0 - x^*||^2 \leq R^2$.
- If additionally, f is μ-strongly convex? (Recall strong convexity implies $f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\mu}{2}||y - x||^2$, for all x, y)

Randomized Coordinate Descent: Convergence Result

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- Define $L = \max_i L_i$ as the Maximum coordinate wise Lipschitz smoothness. Also, assume $||x_0 - x^*||^2 \leq R^2$.
- If additionally, f is μ-strongly convex? (Recall strong convexity implies $f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\mu}{2}||y - x||^2$, for all x, y)
- **Theorem:** If f is coordinate wise Lipschitz Smooth + Strongly Convex, we have:

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

Randomized Coordinate Descent: Improvement at Every iteration!

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x+\gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

Randomized Coordinate Descent: Improvement at Every iteration!

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- Define $L = \max_i L_i$ as the Maximum coordinate wise Lipschitz smoothness. Also, assume $||x_0 - x^*||^2 \leq R^2$.

Randomized Coordinate Descent: Improvement at Every iteration!

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- Define $L = \max_i L_i$ as the Maximum coordinate wise Lipschitz smoothness. Also, assume $||x_0 - x^*||^2 \leq R^2$.
- It holds that:

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}[\nabla f(x_k)]_{i_k}^2$$

Randomized Coordinate Descent: Improvement at Every iteration!

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- Define $L = \max_i L_i$ as the Maximum coordinate wise Lipschitz smoothness. Also, assume $||x_0 - x^*||^2 \leq R^2$.
- It holds that:

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}[\nabla f(x_k)]_{i_k}^2$$

- This means that similar to GD, CD improves the objective value at every iteration!

Importance Sampling

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

Importance Sampling

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x+\gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- Instead of random selection, can we be slightly smarter?

Importance Sampling

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x+\gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- Instead of random selection, can we be slightly smarter?
- Sample i_k with probability: $P[i_k = i] = \frac{L_i}{\sum_i L_i}$, and use step size $1/L_{i_k}$, we can get a slightly better rate of convergence

Importance Sampling

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- Instead of random selection, can we be slightly smarter?
- Sample i_k with probability: $P[i_k = i] = \frac{L_i}{\sum_i L_i}$, and use step size $1/L_{i_k}$, we can get a slightly better rate of convergence
  1. Smooth Functions: $\frac{2d\bar{L}R_0^2}{k}$

Importance Sampling

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- Instead of random selection, can we be slightly smarter?
- Sample i_k with probability: $P[i_k = i] = \frac{L_i}{\sum_i L_i}$, and use step size $1/L_{i_k}$, we can get a slightly better rate of convergence
  1. Smooth Functions: $\frac{2d\bar{L}R_0^2}{k}$
  2. Smooth + Strongly Convex: $(1 - \frac{\mu}{d\bar{L}})^k(f(x_0) - f_*)$

Importance Sampling

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- Instead of random selection, can we be slightly smarter?
- Sample i_k with probability: $P[i_k = i] = \frac{L_i}{\sum_i L_i}$, and use step size $1/L_{i_k}$, we can get a slightly better rate of convergence
  1. Smooth Functions: $\frac{2d\bar{L}R_0^2}{k}$
  2. Smooth + Strongly Convex: $(1 - \frac{\mu}{d\bar{L}})^k(f(x_0) - f_*)$
- In the above, $\bar{L} = \sum_i L_i/d$

Importance Sampling

- Key ingredient: **Coordinate wise Smoothness**:

$$f(x + \gamma e_i) \leq f(x) + \gamma \nabla_i f(x) + \frac{L_i}{2}\gamma^2, \forall x \in \mathbb{R}^d, \forall \gamma \in \mathbb{R}$$

- Instead of random selection, can we be slightly smarter?
- Sample i_k with probability: $P[i_k = i] = \frac{L_i}{\sum_i L_i}$, and use step size $1/L_{i_k}$, we can get a slightly better rate of convergence
  1. Smooth Functions: $\frac{2d\bar{L}R_0^2}{k}$
  2. Smooth + Strongly Convex: $(1 - \frac{\mu}{d\bar{L}})^k (f(x_0) - f_*)$
- In the above, $\bar{L} = \sum_i L_i / d$
- Note that $\bar{L} \leq L$ and can in fact be significantly smaller!

Steepest (or Greedy) Coordinate Descent

- Instead of random or importance sampling based selection, select greedily!

Steepest (or Greedy) Coordinate Descent

- Instead of random or importance sampling based selection, select greedily!
- Select:

$$i_k = \text{argmax}_{i \in [d]} [\nabla_i f(x_k)]$$

Steepest (or Greedy) Coordinate Descent

- Instead of random or importance sampling based selection, select greedily!
- Select:
$$i_k = \text{argmax}_{i \in [d]} [\nabla_i f(x_k)]$$
- Note, this strategy is deterministic instead of random.

Steepest (or Greedy) Coordinate Descent

- Instead of random or importance sampling based selection, select greedily!
- Select:

$$i_k = \mathrm{argmax}_{i \in [d]} [\nabla_i f(x_k)]$$

- Note, this strategy is deterministic instead of random.
- Vanilla analysis gives the same worst case Convergence rate as the randomized case.

Steepest (or Greedy) Coordinate Descent

- Instead of random or importance sampling based selection, select greedily!
- Select:
$$i_k = \text{argmax}_{i \in [d]}[\nabla_i f(x_k)]$$
- Note, this strategy is deterministic instead of random.
- Vanilla analysis gives the same worst case Convergence rate as the randomized case.
- However, if we assume f is strongly convex w.r.t the l_1 norm with parameter $\mu_1 > 0$, we get improved rates of:

Steepest (or Greedy) Coordinate Descent

- Instead of random or importance sampling based selection, select greedily!
- Select:

$$i_k = \text{argmax}_{i \in [d]} [\nabla_i f(x_k)]$$

- Note, this strategy is deterministic instead of random.
- Vanilla analysis gives the same worst case Convergence rate as the randomized case.
- However, if we assume f is strongly convex w.r.t the l_1 norm with parameter $\mu_1 > 0$, we get improved rates of:
  1. Smooth Functions: $\frac{2LR_0^2}{k}$

Steepest (or Greedy) Coordinate Descent

- Instead of random or importance sampling based selection, select greedily!
- Select:

$$i_k = \text{argmax}_{i \in [d]} [\nabla_i f(x_k)]$$

- Note, this strategy is deterministic instead of random.
- Vanilla analysis gives the same worst case Convergence rate as the randomized case.
- However, if we assume f is strongly convex w.r.t the l_1 norm with parameter $\mu_1 > 0$, we get improved rates of:
  1. Smooth Functions: $\frac{2LR_0^2}{k}$
  2. Smooth + Strongly Convex: $(1 - \frac{\sigma}{L})^k (f(x_0) - f_*)$

Exact Coordinate Minimization

- So far, we went over coordinate descent in a way akin to gradient descent, i.e. taking a coordinate step in each coordinate.
- Next, consider exact minimization in each coordinate.

$$
\begin{aligned}
x_1^{(k)} &\in \underset{x_1}{\operatorname{argmin}} \; f\left(x_1, x_2^{(k-1)}, x_3^{(k-1)}, \ldots x_n^{(k-1)}\right) \\
x_2^{(k)} &\in \underset{x_2}{\operatorname{argmin}} \; f\left(x_1^{(k)}, x_2, x_3^{(k-1)}, \ldots x_n^{(k-1)}\right) \\
x_3^{(k)} &\in \underset{x_2}{\operatorname{argmin}} \; f\left(x_1^{(k)}, x_2^{(k)}, x_3, \ldots x_n^{(k-1)}\right) \\
&\ldots \\
x_n^{(k)} &\in \underset{x_2}{\operatorname{argmin}} \; f\left(x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, \ldots x_n\right)
\end{aligned}
$$

for $k = 1, 2, 3, \ldots$

Note: after we solve for $x_i^{(k)}$, we use its new value from then on!

DALLAS

Exact Coordinate Minimization: Linear Regression

Consider linear regression

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2$$

where $y \in \mathbb{R}^n$, and $X \in \mathbb{R}^{n \times p}$ with columns $X_1, \ldots X_p$

Minimizing over β_i, with all β_j, $j \neq i$ fixed:

$$0 = \nabla_i f(\beta) = X_i^T(X\beta - y) = X_i^T(X_i\beta_i + X_{-i}\beta_{-i} - y)$$

i.e., we take

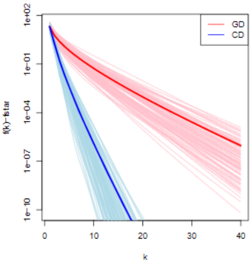
$$\beta_i = \frac{X_i^T(y - X_{-i}\beta_{-i})}{X_i^T X_i}$$

Coordinate descent repeats this update for $i = 1, 2, \ldots, p, 1, 2, \ldots$

LAS

Exact Coordinate Minimization: Linear Regression

Coordinate descent vs gradient descent for linear regression: 100 instances ($n = 100$, $p = 20$)



Is it fair to compare 1 cycle of coordinate descent to 1 iteration of gradient descent? Yes, if we're clever:

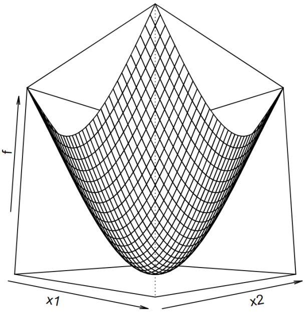
$$\beta_i \leftarrow \frac{X_i^T(y - X_{-i}\beta_{-i})}{X_i^T X_i} = \frac{X_i^T r}{\|X_i\|_2^2} + \beta_i$$

where $r = y - X\beta$. Therefore each coordinate update takes $O(n)$ operations — $O(n)$ to update r, and $O(n)$ to compute $X_i^T r$ — and one cycle requires $O(np)$ operations, just like gradient descent

Optimality Conditions with Coordinate Descent

- Coordinate Descent Family of Algorithms try to make progress every iteration!
- What is the stopping/optimality condition?
- One way to think of this is: If we are a point x such that it is minimum along each coordinate axis! In other words, no coordinate descent algorithm can make progress!
- Does this mean we are at a global minimum?
- Mathematically this means: Does a point x satisfying $f(x + \delta e_i) \geq f(x), \forall \delta, i \Rightarrow f(x) = \min_z f(z)$?

Optimality Conditions: Differentiable Functions

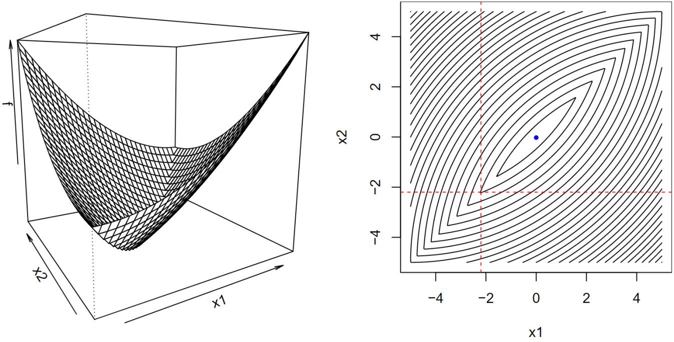


A: Yes! Proof:

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \ldots \frac{\partial f}{\partial x_n}(x) \right) = 0$$

Q: Same question, but for f convex (not differentiable) ... ?

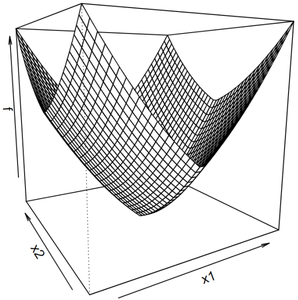
Optimality Conditions: Non Differentiable Functions

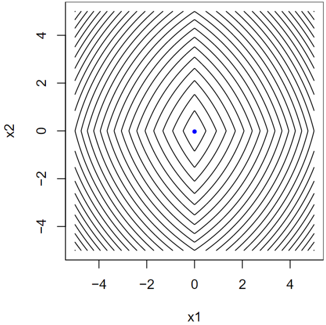


A: No! Look at the above counterexample

Q: Same question again, but now $f(x) = g(x) + \sum_{i=1}^{n} h_i(x_i)$, with g convex, differentiable and each h_i convex ... ? (Nonsmooth part here called separable)

Optimality Conditions: Seperable Functions





A: Yes! Proof: for any y,

$$f(y) - f(x) \geq \nabla g(x)^T (y - x) + \sum_{i=1}^{n} [h_i(y_i) - h_i(x_i)]$$

$$= \sum_{i=1}^{n} \underbrace{[\nabla_i g(x)(y_i - x_i) + h_i(y_i) - h_i(x_i)]}_{\geq 0} \geq 0$$

Coordinate Descent Summary

- Algorithms:

Coordinate Descent Summary

- Algorithms:
  1. Randomized Coordinate Descent

Coordinate Descent Summary

- Algorithms:
  1. Randomized Coordinate Descent
  2. Greedy (Steepest) Coordinate Descent

Coordinate Descent Summary

- Algorithms:
  1. Randomized Coordinate Descent
  2. Greedy (Steepest) Coordinate Descent
  3. Exact Coordinate Minimization

Coordinate Descent Summary

- Algorithms:
  1. Randomized Coordinate Descent
  2. Greedy (Steepest) Coordinate Descent
  3. Exact Coordinate Minimization
- Convergence Results:

Coordinate Descent Summary

- Algorithms:
  1. Randomized Coordinate Descent
  2. Greedy (Steepest) Coordinate Descent
  3. Exact Coordinate Minimization
- Convergence Results:
  1. Randomized Coordinate Descent: Smooth Functions ($\frac{2dLR_0^2}{k}$

Coordinate Descent Summary

- Algorithms:
  1. Randomized Coordinate Descent
  2. Greedy (Steepest) Coordinate Descent
  3. Exact Coordinate Minimization
- Convergence Results:
  1. Randomized Coordinate Descent: Smooth Functions ($\frac{2dLR_0^2}{k}$
  2. Smooth + Strongly Convex: $(1 - \frac{\sigma}{dL})^k (f(x_0) - f_*)$

Coordinate Descent Summary

- Algorithms:
  1. Randomized Coordinate Descent
  2. Greedy (Steepest) Coordinate Descent
  3. Exact Coordinate Minimization
- Convergence Results:
  1. Randomized Coordinate Descent: Smooth Functions ($\frac{2dLR_0^2}{k}$
  2. Smooth + Strongly Convex: $(1 - \frac{\sigma}{dL})^k(f(x_0) - f_*)$
  3. Slight Improvement for Steepest CD: $(1 - \frac{\sigma}{dL})^k(f(x_0) - f_*)$

Coordinate Descent Summary

- Algorithms:
  1. Randomized Coordinate Descent
  2. Greedy (Steepest) Coordinate Descent
  3. Exact Coordinate Minimization
- Convergence Results:
  1. Randomized Coordinate Descent: Smooth Functions ($\frac{2dLR_0^2}{k}$
  2. Smooth + Strongly Convex: $(1 - \frac{\sigma}{dL})^k(f(x_0) - f_*)$
  3. Slight Improvement for Steepest CD: $(1 - \frac{\sigma}{dL})^k(f(x_0) - f_*)$
  4. Similar bound for exact coordinate minimization. Method of choice if the coordinate wise minimization can be solved exactly!

Coordinate Descent Summary

- Algorithms:
  1. Randomized Coordinate Descent
  2. Greedy (Steepest) Coordinate Descent
  3. Exact Coordinate Minimization
- Convergence Results:
  1. Randomized Coordinate Descent: Smooth Functions ($\frac{2dLR_0^2}{k}$
  2. Smooth + Strongly Convex: $(1 - \frac{\sigma}{dL})^k (f(x_0) - f_*)$
  3. Slight Improvement for Steepest CD: $(1 - \frac{\sigma}{dL})^k (f(x_0) - f_*)$
  4. Similar bound for exact coordinate minimization. Method of choice if the coordinate wise minimization can be solved exactly!
  5. Above bounds can also be extended if the non-differentiability is seperable (e.g. L1 regularization).

Coordinate Descent Summary

- Algorithms:
  1. Randomized Coordinate Descent
  2. Greedy (Steepest) Coordinate Descent
  3. Exact Coordinate Minimization
- Convergence Results:
  1. Randomized Coordinate Descent: Smooth Functions ($\frac{2dLR_0^2}{k}$
  2. Smooth + Strongly Convex: $(1 - \frac{\sigma}{dL})^k(f(x_0) - f_*)$
  3. Slight Improvement for Steepest CD: $(1 - \frac{\sigma}{dL})^k(f(x_0) - f_*)$
  4. Similar bound for exact coordinate minimization. Method of choice if the coordinate wise minimization can be solved exactly!
  5. Above bounds can also be extended if the non-differentiability is seperable (e.g. L1 regularization).
- Extensions: Accelerated Coordinate Minimization also possible!