(Sub)Gradient Descent: Generalization of Gradient Descent

Given a convex function $f\colon \mathbb{R}^n \to \mathbb{R}$, not necessarily differentiable. Subgradient method is just like gradient descent, but replacing gradients with subgradients. *I.e.*, initialize $x^{(0)}$, then repeat

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - t^k \cdot \mathbf{h}^{(k-1)}, k = 1, 2, 3, \cdots$$

where $\mathbf{h}^{(k-1)}$ is **any** subgradient of f at $\mathbf{x}^{(k-1)}$. We keep track of best iterate $\mathbf{x}_{best}^{k}$ among $\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(k)}$:

$$f(\mathbf{x}_{best}^{(k)}) = \min_{i=1,\cdots,k} f(\mathbf{x}^{(i)})$$

To update each $\mathbf{x}^{(i)}$, there are basically two ways to select the step size:

- Fixed step size: $t^k = t$ for all $k = 1, 2, 3 \cdots$
- Diminishing step size: choose t^k to satisfy

$$\lim_{k\to\infty} (t^k) = 0 \ , \ \sum_{k=1}^{\infty} t^k = \infty$$

Eg: t^k = 1/k
t^k = 1/sqrt{k}

Subgradient Algorithm: Convergence analysis

Given the convex function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies:

- f is Lipschitz continuous with constant $l > 0$,

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq l||\mathbf{x} - \mathbf{y}|| \text{ for all } \mathbf{x}, \mathbf{y}$$

- $||\mathbf{x}^{(1)} - \mathbf{x}^*|| \leq R$ which means it is bounded All iterates lie within R

Theorem

For a fixed step size t, subgradient method satisfies

$$\lim_{k\to\infty} f(x_{best}^{(k)}) \leq f(\mathbf{x}^*) + \frac{l^2 t}{2}$$

For diminishing step size such as $t^k = O\left(\frac{1}{\sqrt{k}}\right)$,

$$f(x_{best}^{(k)}) \leq f(\mathbf{x}^*) + O\left(\frac{1}{\sqrt{k}}\right)$$

Subgradient Descent: Convergence Analysis (contd.)

Proof:

$$
\begin{aligned}
||\mathbf{x}^{(k+1)} - \mathbf{x}^*||^2 &= ||\mathbf{x}^{(k)} - t^k\mathbf{h}^{(k)} - \mathbf{x}^*||^2 \\
&= ||\mathbf{x}^{(k)} - \mathbf{x}^*||^2 - 2t^k(\mathbf{h}^{(k)})^T(\mathbf{x}^{(k)} - \mathbf{x}^*) + (t^k)^2||\mathbf{h}^{(k)}||^2
\end{aligned}
$$

By definition of the subgradient method, we have

$$
\begin{aligned}
f(\mathbf{x}^*) &\geq f(\mathbf{x}^{(k)}) + (\mathbf{h}^{(k)})^T(\mathbf{x}^* - \mathbf{x}^{(k)}) \\
-(\mathbf{h}^{(k)})^T(\mathbf{x}^* - \mathbf{x}^{(k)}) &\leq -(f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*))
\end{aligned}
$$

Using this inequality, for $k, k-1, \ldots i, i-1, \ldots 0$ we have

Subgradient Descent: Convergence Analysis (contd.)

Proof:

$$\begin{aligned} \|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|^2 &= \|\mathbf{x}^{(k)} - t^k\mathbf{h}^{(k)} - \mathbf{x}^*\|^2 \\ &= \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2 - 2t^k(\mathbf{h}^{(k)})^T(\mathbf{x}^{(k)} - \mathbf{x}^*) + (t^k)^2\|\mathbf{h}^{(k)}\|^2 \end{aligned}$$

By definition of the subgradient method, we have

$$f(\mathbf{x}^*) \geq f(\mathbf{x}^{(k)}) + (\mathbf{h}^{(k)})^T(\mathbf{x}^* - \mathbf{x}^{(k)})$$
$$-(\mathbf{h}^{(k)})^T(\mathbf{x}^* - \mathbf{x}^{(k)}) \leq -(f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*))$$

Using this inequality, for $k, k-1, \ldots i, i-1, \ldots 0$ we have

$$\begin{aligned} \|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|^2 &\leq \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2 - 2t^k(f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*)) + (t^k)^2\|\mathbf{h}^{(k)}\|^2 \\ &\leq \|\mathbf{x}^{(1)} - \mathbf{x}^*\|^2 - 2\sum_{i=1}^{k} t^i(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)) + \sum_{i=1}^{k}(t^i)^2\|\mathbf{h}^{(i)}\|^2 \end{aligned}$$

Subgradient Descent: Convergence Analysis (contd.)

And since this is lower bounded by 0, we have

$$0 \leq ||\mathbf{x}^{(k+1)} - \mathbf{x}^*||^2 \leq R^2 - 2\sum_{i=1}^{k} t^i(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)) + \sum_{i=1}^{k}(t^i)^2 l^2$$

$$\Rightarrow\ 2\sum_{i=1}^{k} t^i(f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)) \leq R^2 + \sum_{i=1}^{k}(t^i)^2 l^2$$

$$\Rightarrow\ 2(\sum_{i=1}^{k} t^i)(f(\mathbf{x}^{(k)}_{best}) - f(\mathbf{x}^*)) \leq R^2 + \sum_{i=1}^{k}(t^i)^2 l^2$$

Using Lipschitz continuity of f, ||h||^2 <= l^2
Using the fact that best iterate is always maintained, f(x^i) >= f(x^k_be
Using the fact that ||x - x*||^2 <= R^2

Subgradient Descent: Convergence Analysis (contd.)

For a constant step size $t^i = t$:

$$\frac{R^2 + l^2 t^2 k}{2tk} \rightarrow \frac{l^2 t}{2}, \text{ as } k \rightarrow \infty,$$

and for diminishing step size, we have:

Subgradient Descent: Convergence Analysis (contd.)

For a constant step size $t^i = t$:

$$\frac{R^2 + l^2 t^2 k}{2tk} \to \frac{l^2 t}{2}, \text{ as } k \to \infty,$$

and for diminishing step size, we have:

$$\sum_{i=0}^{k} (t^i)^2 \leq 0, \sum_{i=0}^{k} t^i = \infty$$

therefore,

$$\frac{R^2 + l^2 \sum_{i=0}^{k} (t^i)^2}{2 \sum_{i=0}^{k} t^i} \to 0, \text{ as } k \to \infty,$$

$\square$

Subgradient Descent: Convergence Analysis (contd.)

Consider taking $t^i = R/(l\sqrt{k})$, for all $i = 1, ..., k$. Then we can obtain the following tendency:

Subgradient Descent: Convergence Analysis (contd.)

Consider taking $t^i = R/(l\sqrt{k})$, for all $i = 1, ..., k$. Then we can obtain the following tendency:

$$\frac{R^2 + l^2 \sum_{i=0}^{k} (t^i)^2}{2 \sum_{i=0}^{k} t^i} = \frac{Rl}{\sqrt{k}}.$$

That is, subgradient method has convergence rate of $O(\frac{1}{\sqrt{k}})$, and to get $f(x_{best}^{(k)}) - f(x^*) \leq \epsilon$, needs $O(\frac{1}{\epsilon^2})$ iterations.

This is a much worse convergence rate than even $O\left(\frac{1}{k}\right)$ obtained for gradient descent under Lipschitz continuity alone.

We again see why Lipschitz continuity of function is a somewhat weaker assumption that Lipschitz continuity of the gradient

Optimization: Subgradient Descent and Constrained Optimization

Instructor: Prof. Ganesh Ramakrishnan

We will evolve specific and more efficient, better motivated special cases (or even variants) of subgradient descent for constraint optimization, mostly using the fact that the objective function is often differentiable and sometimes has Lipschitz continuous gradient

Constrained Optimization and Subgradient Descent

Constrained Optimization

- Consider the objective

$$\min f(x)$$
$$\text{s.t. } g_i(x) \leq 0, \forall i$$

- Recall: Indicator function for $g_i(x)$

$$I_{g_i}(x) = \begin{cases} 0, & \text{if } g_i(x) \leq 0 \\ \infty, & \text{otherwise} \end{cases}$$

  - We have shown that this is convex if each $g_i(\mathbf{x})$ is convex.
- Option 1: Minimize sum of f(x) and I_g_i(x)

Constrained Optimization

- Consider the objective

$$\min f(x)$$
$$\text{s.t. } g_i(x) \leq 0, \forall i$$

- Recall: Indicator function for $g_i(x)$

$$I_{g_i}(x) = \begin{cases} 0, & \text{if } g_i(x) \leq 0 \\ \infty, & \text{otherwise} \end{cases}$$

  - We have shown that this is convex if each $g_i(\mathbf{x})$ is convex.
- Option 1: Use subgradient descent to solve this optimization
- Option 2: Write down smooth (differentiable) approximation functions for the Indicator...Often bringing in additional advantages such as strong convexity (even if f(x) was not strongly convex

Constrained Optimization

- Consider the objective

$$\min f(x)$$

$$\text{s.t. } g_i(x) \leq 0, \forall i$$

- Recall: Indicator function for $g_i(x)$

$$I_{g_i}(x) = \begin{cases} 0, & \text{if } g_i(x) \leq 0 \\ \infty, & \text{otherwise} \end{cases}$$

  - We have shown that this is convex if each $g_i(\mathbf{x})$ is convex.
- Option 1: Use subgradient descent to solve this optimization
- Option 2: Barrier Method (approximate $I_{g_i}(\mathbf{x})$ using some differentiable function), Augmented Lagrangian, ADMM, *etc.* Later...

Option 1: (Sub)Gradient Descent with Sum of indicators

- Convert our objective to the following unconstrained optimization problem
- Each $C_i = \{\mathbf{x} \mid g_i(\mathbf{x}) \leq 0\}$ is convex if $g_i(\mathbf{x})$ is convex.
- We take

$$\min_{\mathbf{x}} F(\mathbf{x}) = \min_{\mathbf{x}} f(\mathbf{x}) + \sum_i I_{C_i}(\mathbf{x})$$

- Recap a subgradient of F:

gradient of f(x) + linear combination of elements
of normal cones to each C_i

Option 1: (Sub)Gradient Descent with Sum of indicators

- Convert our objective to the following unconstrained optimization problem
- Each $C_i = \{\mathbf{x} \mid g_i(\mathbf{x}) \leq 0\}$ is convex if $g_i(\mathbf{x})$ is convex.
- We take

$$\min_{\mathbf{x}} F(\mathbf{x}) = \min_{\mathbf{x}} f(\mathbf{x}) + \sum_i I_{C_i}(\mathbf{x})$$

- Recap a subgradient of F: $\mathbf{h}_F(\mathbf{x}) = \underline{\mathbf{h}_f(\mathbf{x})} + \sum_i \mathbf{h}_{I_{C_i}}(\mathbf{x})$. Recall that

Gradient...

x-gradient f(x) minimizes the quadratic approximation of the function f around x

Option 1: (Sub)Gradient Descent with Sum of indicators

- Convert our objective to the following unconstrained optimization problem
- Each $C_i = \{\mathbf{x} \mid g_i(\mathbf{x}) \leq 0\}$ is convex if $g_i(\mathbf{x})$ is convex.
- We take

$$\text{min}_{\mathbf{x}} F(\mathbf{x}) = \min_{\mathbf{x}} f(\mathbf{x}) + \sum_i I_{C_i}(\mathbf{x})$$

- Recap a subgradient of F: $\mathbf{h}_F(\mathbf{x}) = \mathbf{h}_f(\mathbf{x}) + \sum_i \mathbf{h}_{I_{C_i}}(\mathbf{x})$. Recall that
  - $\mathbf{h}_f(\mathbf{x}) = \nabla f(\mathbf{x})$ if $f(\mathbf{x})$ is differentiable. Also, $-\nabla f(\mathbf{x})$ at $\mathbf{x}^k$ optimizes

Option 1: (Sub)Gradient Descent with Sum of indicators

- Convert our objective to the following unconstrained optimization problem
- Each $C_i = \{\mathbf{x} \mid g_i(\mathbf{x}) \leq 0\}$ is convex if $g_i(\mathbf{x})$ is convex.
- We take

$$\min_{\mathbf{x}} F(\mathbf{x}) = \min_{\mathbf{x}} f(\mathbf{x}) + \sum_i I_{C_i}(\mathbf{x})$$

- Recap a subgradient of F: $\mathbf{h}_F(\mathbf{x}) = \mathbf{h}_f(\mathbf{x}) + \sum_i \mathbf{h}_{I_{C_i}}(\mathbf{x})$. Recall that
  - $\mathbf{h}_f(\mathbf{x}) = \nabla f(\mathbf{x})$ if $f(\mathbf{x})$ is differentiable. Also, $-\nabla f(\mathbf{x})$ at $\mathbf{x}^k$ optimizes the first order approximation for $f(\mathbf{x})$ around $\mathbf{x}^k$: $-\nabla f(\mathbf{x}) = \underset{\mathbf{h}}{\operatorname{argmin}}\, f(\mathbf{x}^k) + \nabla^T f(\mathbf{x}^k)\mathbf{h} + \frac{1}{2}||\mathbf{h}||^2$ **Leads to Mirror Descent etc.** All examples hereafter are special case of MD
  - $\mathbf{h}_{I_{C_i}}(x)$ is $d \in \mathbf{R}^n$ s.t. $d^\top x \geq d^\top y$, $\forall y \in C_i$. Also, $\mathbf{h}_{I_{C_i}}(x) = 0$ if $\mathbf{x}$ is in the interior of C_i, and has other solutions if $\mathbf{x}$ is on the boundary. **Leads to KKT conditions and Dual Ascent etc.**

Changing the notion of distance from squared (to say 1 norm) can yield different approximations, different notion of strong convexity...

Option 1: Generalized Gradient Descent

- Consider the following sum of a differentiable function $f(\mathbf{x})$ and a nondifferentiable function $c(\mathbf{x})$ (an example being $\sum_i I_{C_i}(\mathbf{x})$)
- We take

Recap our discussion on subgradient for proxim

$$\min_{\mathbf{x}} F(\mathbf{x}) = \min_{\mathbf{x}} \boxed{f(\mathbf{x}) + c(\mathbf{x})}$$

- Like gradient descent, consider the first order approximation for $f(\mathbf{x})$ around $\mathbf{x}^k$ leaving $c(\mathbf{x})$ alone to obtain the next iterate $\mathbf{x}^{k+1}$:

$$\mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{x}} \boxed{\cancel{f(\mathbf{x}^k)} + \nabla^T f(\mathbf{x}^k)(\mathbf{x} - \mathbf{x}^k) + \frac{1}{2t}||\mathbf{x} - \mathbf{x}^k||^2 + c(\mathbf{x})}$$

- Deleting $\cancel{f(\mathbf{x}^k)}$ from the objective and adding $\frac{t}{2}||\nabla f(\mathbf{x}^k)||^2$ to the objective (without any loss) to complete squares, we obtain $\mathbf{x}^{k+1}$ as:

Find next iterate as close as possible to grad desc iterate of f, while also minimizaing c

$$\mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{x}} \boxed{\frac{1}{2t}||\mathbf{x} - (\mathbf{x}^k - t\nabla f(\mathbf{x}^k))||^2 + c(\mathbf{x})}$$

- In general, such a step is called a *proximal* step

$$\mathbf{x}^{k+1} = prox_t\left(\mathbf{x}^k - t\nabla f(\mathbf{x}^k)\right) = \operatorname{argmin}_{\mathbf{x}} \frac{1}{2t}||\mathbf{x} - (\mathbf{x}^k - t\nabla f(\mathbf{x}^k))||^2 + c(\mathbf{x})$$

Option 1: Generalized Gradient Descent

- Interesting because in many settings, $prox_t(\mathbf{z})$ can be computed efficiently

$$prox_t(\mathbf{z}) = \underset{\mathbf{x}}{\text{argmin}} \frac{1}{2t}||\mathbf{x} - \mathbf{z}||^2 + c(\mathbf{x})$$

- Illustration on Lasso: $\mathbf{x}^* = \underset{\mathbf{x}}{\text{argmin}} ||A\mathbf{x} - \mathbf{y}||^2 + \|\mathbf{x}\|_1$. You can successively use $\mathbf{z} = \mathbf{x}^k - t\nabla f(\mathbf{x}^k)$.
-
- H/w: Build upon midsem solution to write the proximal step for Lasso in terms of z (no need of x^k)
-
-

Illustration on Lasso

Illustration on Lasso

Option 1: Generalized Gradient Descent

- Recall

$$prox_t(\mathbf{z}) = \underset{\mathbf{x}}{\operatorname{argmin}} \frac{1}{2t}||\mathbf{x} - \mathbf{z}||^2 + c(\mathbf{x})$$

1. Gradient Descent: $c(\mathbf{x}) = 0$
2. Projected Gradient Descent: $c(\mathbf{x}) = \sum_i I_{C_i}(\mathbf{x})$
3. Proximal Minimization: $f(\mathbf{x}) = 0$ You are trying to find solution to a bunch of constraints

We will discuss these specific cases after a short discussion on convergence

[7] Else we just treat this as another minimization problem and obtain an approximate solution. Practical convergence rate can be very slow. Exceptions are partial proximation minimization [Bertsekas and Tseng '94]

Option 1: Generalized Gradient Descent

- Recall

$$prox_t(\mathbf{z}) = \operatorname*{argmin}_{\mathbf{x}} \frac{1}{2t}||\mathbf{x}-\mathbf{z}||^2 + c(\mathbf{x})$$

1. Gradient Descent: $c(\mathbf{x}) = 0$
2. Projected Gradient Descent: $c(\mathbf{x}) = \sum_i I_{C_i}(\mathbf{x})$
3. Proximal Minimization: $f(\mathbf{x}) = 0$

We will discuss these specific cases after a short discussion on convergence

- Convergence: If $f(\mathbf{x})$ is convex, differentiable, and ∇f is Lipschitz continuous with constant $L > 0$ AND $c(\mathbf{x})$ **is convex and** $prox_t(\mathbf{z})$ **can be solved exactly**[7] then

f(x^k) - f(x^*) <= ||x^k - x^0||^2/2tk

[7] Else we just treat this as another minimization problem and obtain an approximate solution. Practical convergence rate can be very slow. Exceptions are partial proximation minimization [Bertsekas and Tseng '94]

Option 1: Generalized Gradient Descent

- Recall

$$prox_t(\mathbf{z}) = \operatorname*{argmin}_{\mathbf{x}} \frac{1}{2t}||\mathbf{x}-\mathbf{z}||^2 + c(\mathbf{x})$$

1. Gradient Descent: $c(\mathbf{x}) = 0$
2. Projected Gradient Descent: $c(\mathbf{x}) = \sum_i I_{C_i}(\mathbf{x})$
3. Proximal Minimization: $f(\mathbf{x}) = 0$

We will discuss these specific cases after a short discussion on convergence

- Convergence: If $f(\mathbf{x})$ is convex, differentiable, and ∇f is Lipschitz continuous with constant $L > 0$ AND $c(\mathbf{x})$ **is convex and** $prox_t(\mathbf{z})$ **can be solved exactly**[7] then convergence result (and proof) is similar to that for gradient descent

$$f(x^k) - f(x^*) \le \frac{1}{k}\sum_{i=1}^{k}\left(f(x^i) - f(x^*)\right) \le \frac{\left\|x^{(0)} - x^*\right\|^2}{2tk}$$

[7] Else we just treat this as another minimization problem and obtain an approximate solution. Practical convergence rate can be very slow. Exceptions are partial proximation minimization [Bertsekas and Tseng '94]

Convergence Rate: Generalized Gradient Descent vs. Subgradient Descent

- Recap: For Subgraident Descent: The subgradient method has convergence rate $O(1/\sqrt{k})$; to get $f(\mathbf{x}_{best}^{(k)}) - f(\mathbf{x}^*) \leq \epsilon$, we need $O(1/\sqrt{\epsilon^2})$ iterations.
This is actually the best we can do; e.g., we can't do better than $O(1/\sqrt{k})$.

Proved by Nesterov

Convergence Rate: Generalized Gradient Descent vs. Subgradient Descent

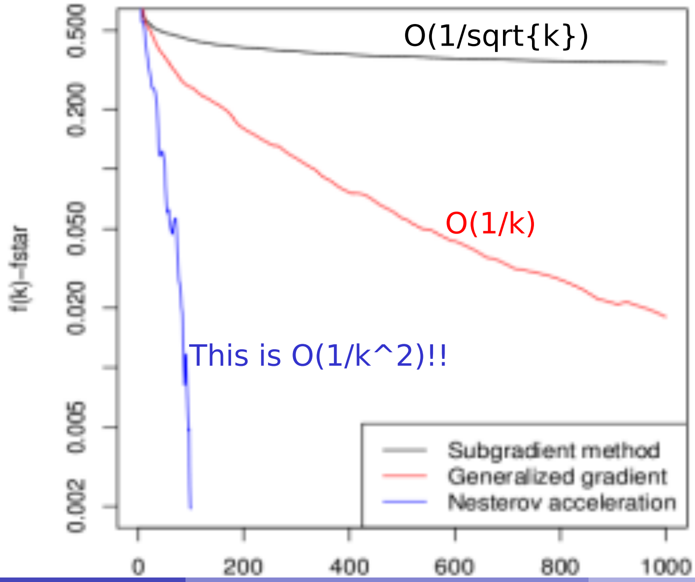
- Recap: For Subgraident Descent: The subgradient method has convergence rate $O(1/\sqrt{k})$; to get $f(\mathbf{x}^{(k)}_{best}) - f(\mathbf{x}^*) \leq \epsilon$, we need $O(1/\sqrt{\epsilon^2})$ iterations.
  This is actually the best we can do; e.g., we can't do better than $O(1/\sqrt{k})$.
- For generalized Gradient Descent: If $f(x)$ is convex, differentiable, and ∇f is Lipschitz continuous with constant $L > 0$ AND $c(x)$ is convex and $prox_t(x)$ can be solved exactly then convergence result (and proof) is similar to that for gradient descent

$$f(x^k) - f(x^*) \leq \frac{1}{k}\sum_{i=1}^{k}\left(f(x^i) - f(x^*)\right) \leq \frac{\left\|x^{(0)} - x^*\right\|^2}{2tk}$$

Better convergence ($O(1/k)$) because of assuming (i) Differentiability of $f(\mathbf{x})$ and (ii) Lipschitz continuity of $\nabla f(\mathbf{x})$.
Can we do even better without strong convexity (not possible for $c(\mathbf{x})$?

Yes. You cannot get Q-linear as in strong convexity. But O(1/k^2) possible

(Nesterov) Accelerated Generalized Gradient Descent



(Nesterov) Accelerated Generalized Gradient Descent

The problem is:

$$\min_{x \in \mathbb{R}^n} f(\mathbf{x}) + c(\mathbf{x})$$

where $f(\mathbf{x})$ is convex and differentiable, $c(\mathbf{x})$ is convex and not necessarily differentiable.

- Initialize $\mathbf{x}_u^{(0)} \in \mathbb{R}^n$
- repeat for $k = 1, 2, 3, \ldots$

Momentum is used to somewhat capture second order moments/curvature

$$\mathbf{y} = \mathbf{x}^{(k-1)} + \frac{k-2}{k+1}(\mathbf{x}^{(k-1)} - \mathbf{x}^{(k-2)})$$

$$\mathbf{x}^{(k)} = \text{prox}_{t^k}(\mathbf{y} - t^k \nabla f(\mathbf{y}))$$

Prox not on x^k-1 but on x^k-1 averaged with x^k-2

Or Equivalently,

$$\mathbf{y} = (1 - \theta_k)\mathbf{x}^{(k-1)} + \theta_k \mathbf{x}_u^{(k-1)}$$

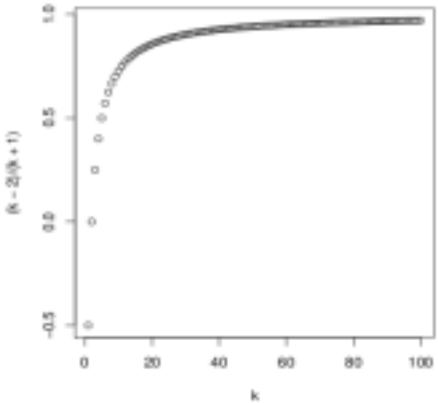
$$\mathbf{x}^k = \text{prox}_{t^k}(\mathbf{y} - t^k \nabla f(\mathbf{y}))$$

$$\mathbf{u}^{(k)} = \mathbf{x}^{(k-1)} + \frac{1}{\theta_k}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})$$

where $\theta = 2/(k+1)$.

(Nesterov) Accelerated Generalized Gradient Descent

1. First step $k = 1$ is just usual generalized gradient update: $\mathbf{x}^{(1)} = \text{prox}_{t^1}\left(\mathbf{x}^{(0)} - t^1 \nabla f(\mathbf{x}^{(0)})\right)$
2. Thereafter, the method carries some "momentum" from previous iterations
3. $c(\mathbf{x}) = 0$ gives accelerated gradient method
4. The method accelerates more towards the end of iterations



(Nesterov) Accelerated Generalized Gradient Descent

Examples showing the performance of accelerated gradient descent compared with usual gradient descent.

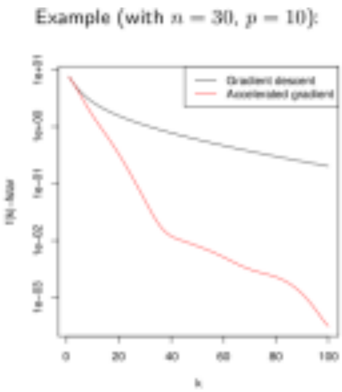


Figure 12: Example 1: Performance of accelerated gradient descent compared with usual gradient descent

(Nesterov) Accelerated Generalized Gradient Descent: Convergence

Minimize $f(\mathbf{x}) = f(\mathbf{x}) + c(\mathbf{x})$ assuming that:
f is convex, differentiable, ∇f is Lipschitz with constant $L > 0$, and
c is convex, the prox function can be evaluated.

Theorem

Accelerated generalized gradient method with fixed step size $t \leq 1/L$ satisfies:

$$f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) \leq \frac{2||x^{(0)} - x^*||^2}{t(k+1)^2}$$

Accelerated generalized gradient method can achieve the optimal $O(1/k^2)$ rate for first-order method, or equivalently, if we want to get $f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) \leq \epsilon$, we only need $O(1/\sqrt{\epsilon})$ iterations. Now we prove this theorem.