Lecture 07

<u>Gradient Descent</u>

Gradient descent

Consider unconstrained, smooth convex optimization

$$\min_x \; f(x)$$

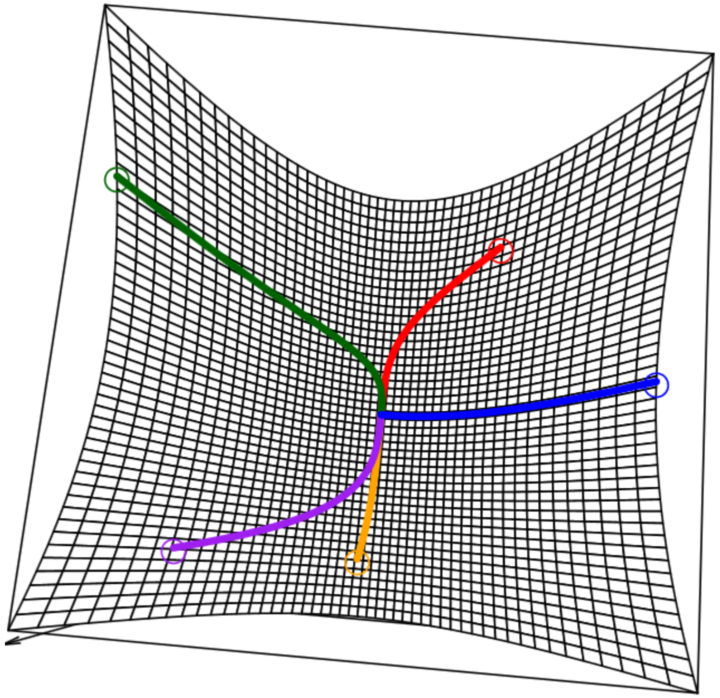
i.e., f is convex and differentiable with $\mathrm{dom}(f) = \mathbb{R}^n$. Denote the optimal criterion value by $f^\star = \min_x \; f(x)$, and a solution by $x^\star$

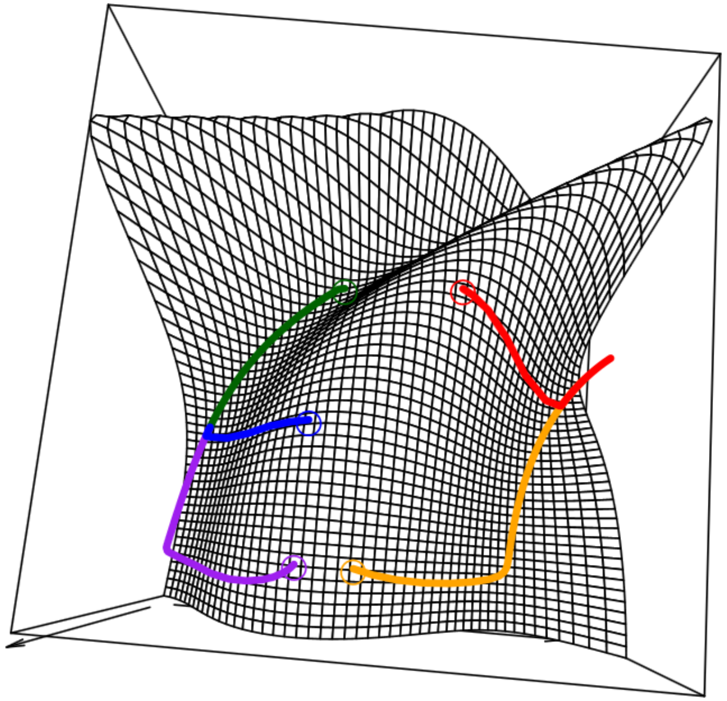
Gradient descent: choose initial point $x^{(0)} \in \mathbb{R}^n$, repeat:

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

Stop at some point

- Think about gradient descent as repeatedly going downhill.
- The negative gradient is going in the direction that decreases the optimization criterion.
- Thus, we will stop at some point close to the minimum solution independent on the starting point. This is valid only for convex functions.
- In non-convex functions, depending on the starting point different local minima could be achieved.





Gradient descent interpretation

- we can interpret gradient descent via a quadratic approximation.

- Suppose we are at point x, and we make a second order Taylor expansion of function $f(y)$.

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

- Replacing, $\nabla^2 f(x) = \frac{1}{t} I$ and thus assuming a proximity term to x equal to $\frac{1}{2t}\|y - x\|_2^2$ with weight $\frac{1}{2t}$, and a linear approximation to f as $f(x) + \nabla f(x)^T \cdot (y - x)$, we have:

$$f(y) \approx f(x) + \nabla f(x)^T \cdot (y - x) + \frac{1}{2t}\|y - x\|_2^2$$

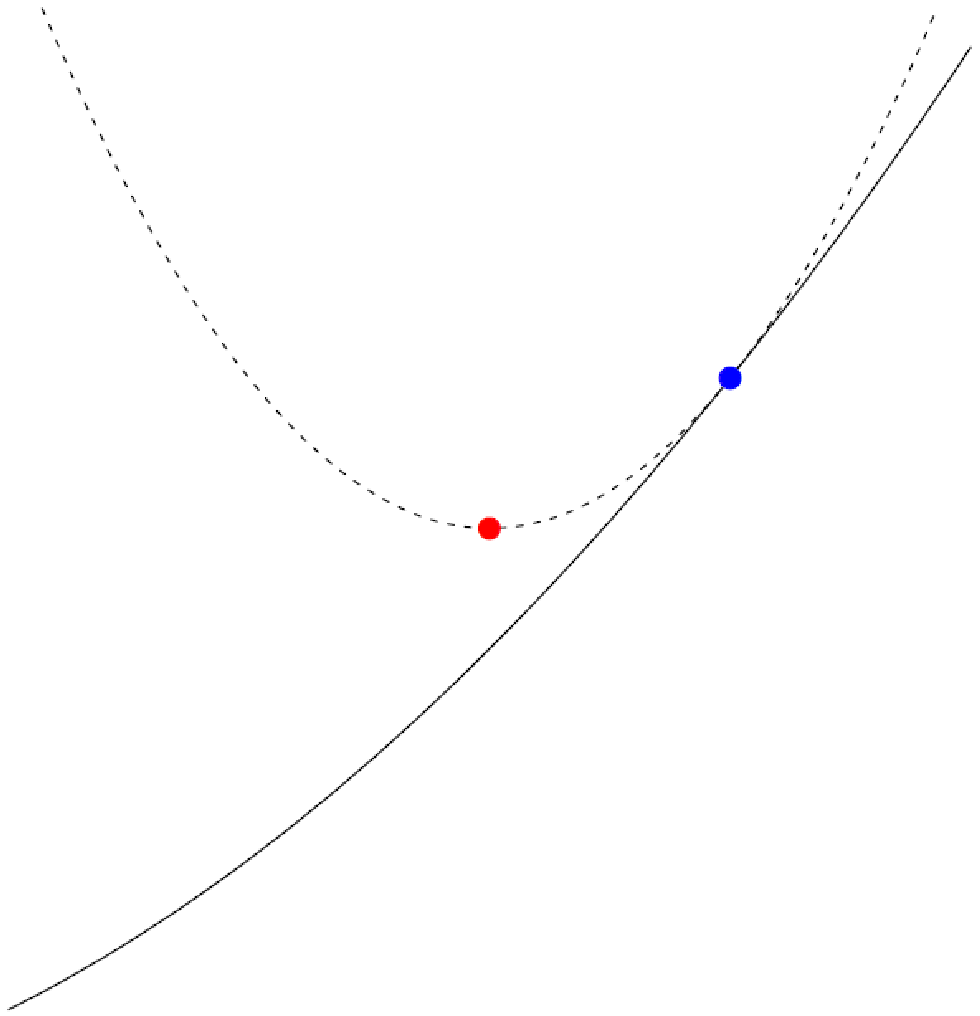
Gradient descent interpretation

- Gradient descent will choose the next point $y = x^+$ to minimize the quadratic approximation by taking the gradient of $f(y)$ equal to zero:

$$x^+ = \underset{y}{\operatorname{argmin}} \quad f(x) + \nabla f(x)^T \cdot (y - x) + \frac{1}{2t} \|y - x\|_2^2$$

$$x^+ = x - t\nabla f(x)$$

- Depending on how close the next step should be to the current state x depends on weight $\frac{1}{2t}$ of the proximity term.
  - If t is small, the weight of the proximity term is large and steps will be small.

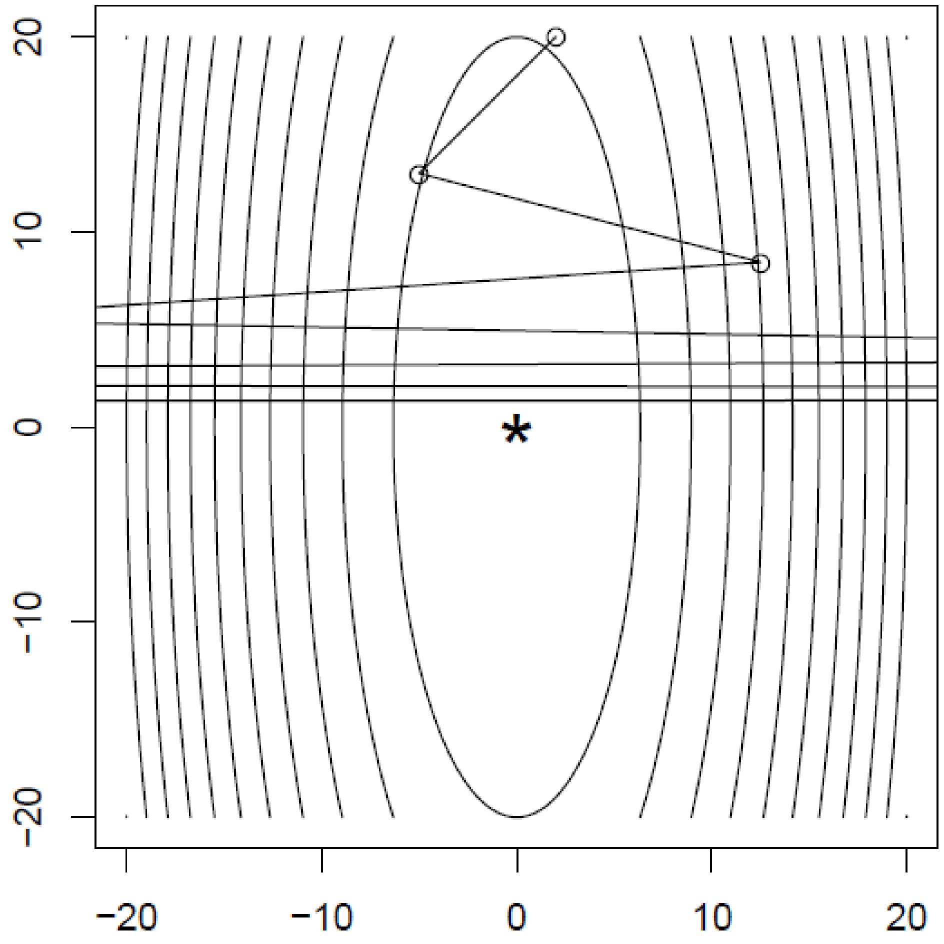


Blue point is x, red point is

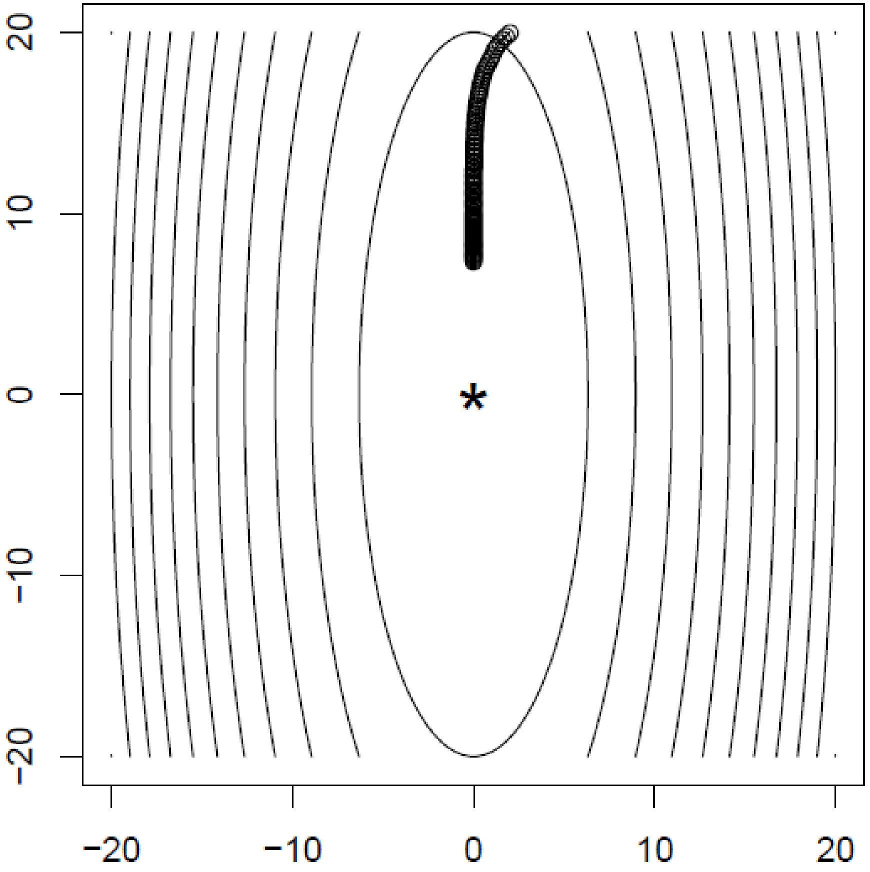
$$x^+ = \underset{y}{\operatorname{argmin}}\ f(x) + \nabla f(x)^T (y - x) + \frac{1}{2t}\|y - x\|_2^2$$

Fixed step size

Simply take $t_k = t$ for all $k = 1, 2, 3, \ldots$, can diverge if t is too big.
Consider $f(x) = (10x_1^2 + x_2^2)/2$, gradient descent after 8 steps:

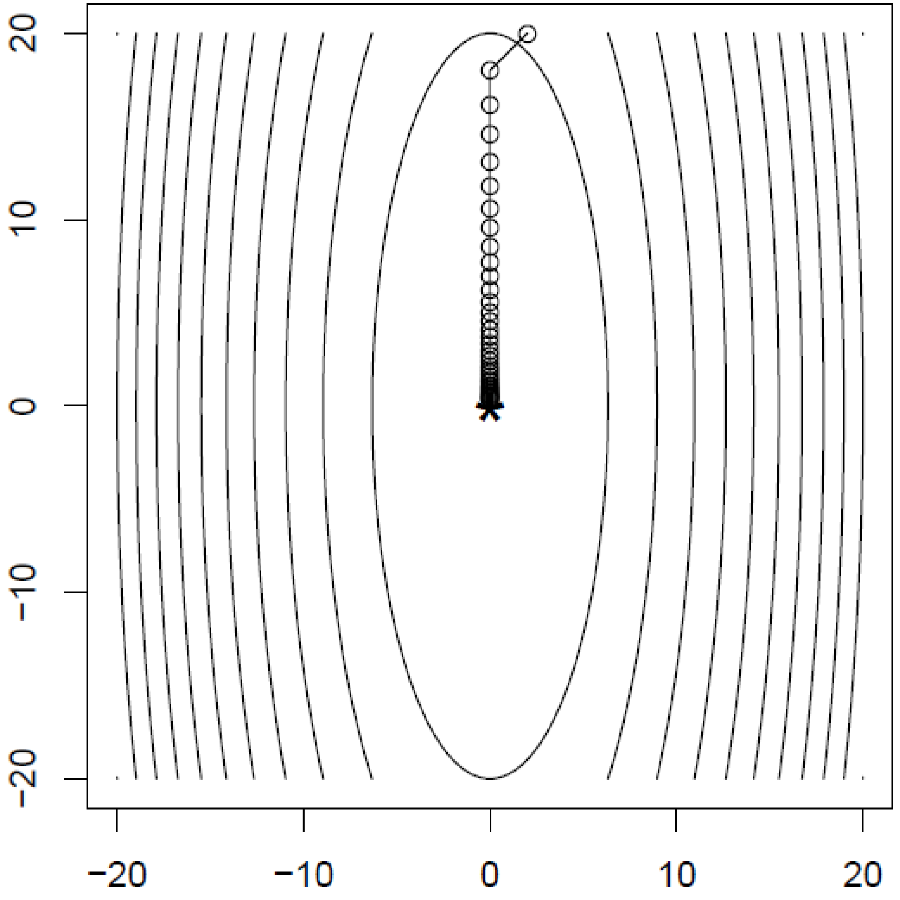


Can be slow if t is too small. Same example, gradient descent after 100 steps:



- As updates get closer to the minimum, the effective step $t\nabla f(x)$ gets small as the gradient $\nabla f(x)$ approaches zero and thus step direction will shrink by default and slowed down the process.

Converges nicely when t is "just right". Same example, gradient descent after 40 steps:



Convergence analysis later will give us a precise idea of "just right"

Exact line search

Could choose step to do the best we can along direction of negative gradient, called exact line search:

$$t = \underset{s \geq 0}{\operatorname{argmin}} \ f(x - s\nabla f(x))$$

Usually not possible to do this minimization exactly

Approximations to exact line search are often not as efficient as backtracking, and it's usually not worth it

Backtracking line search

One way to adaptively choose the step size is to use backtracking line search:

- First fix parameters $0 < \beta < 1$ and $0 < \alpha \leq 1/2$
- At each iteration, start with $t = t_{\text{init}}$, and while

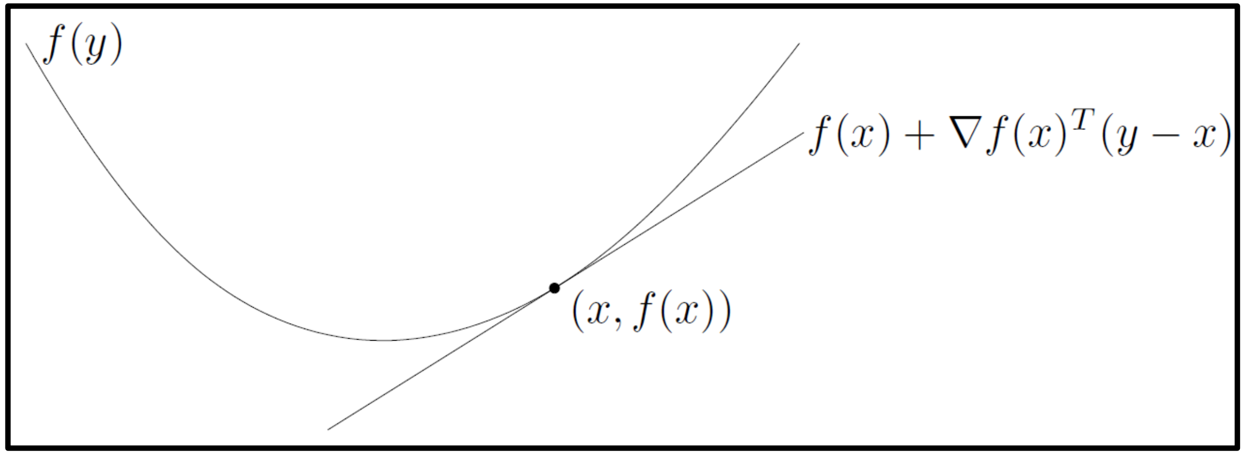
$$f(x - t\nabla f(x)) > f(x) - \alpha t \|\nabla f(x)\|_2^2$$

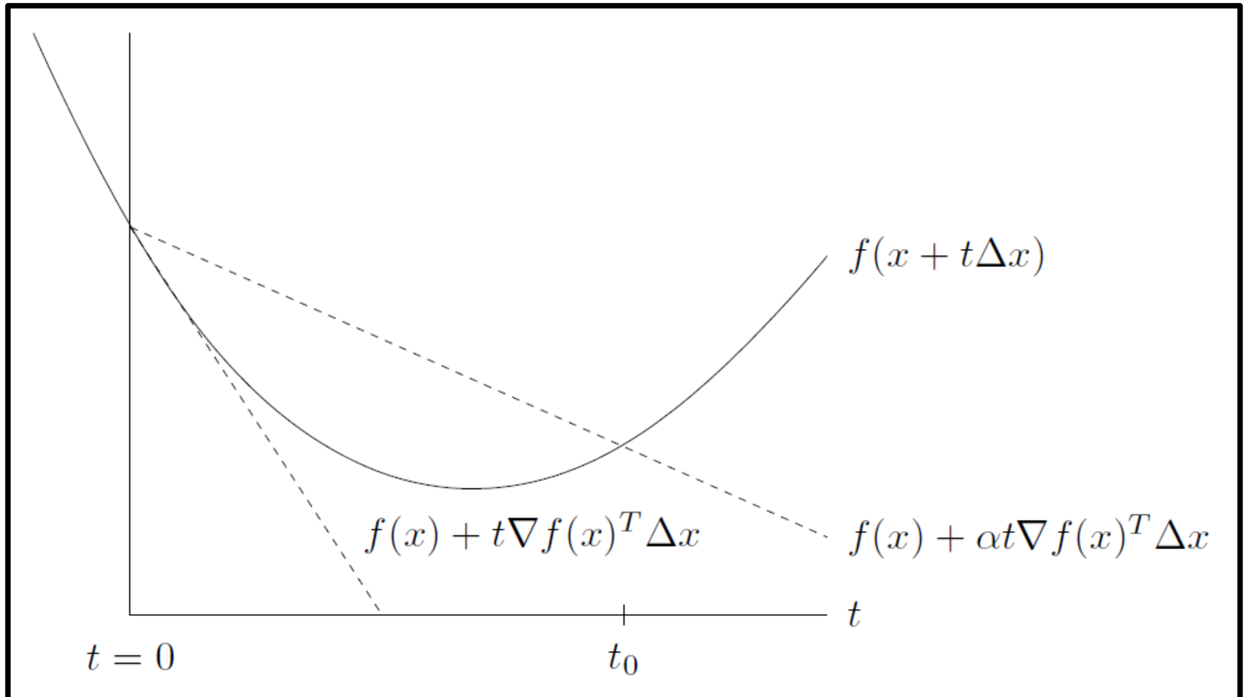
  shrink $t = \beta t$. Else perform gradient descent update

$$x^+ = x - t\nabla f(x)$$

Simple and tends to work well in practice (further simplification: just take $\alpha = 1/2$)

Backtracking interpretation





For $\Delta x = -\nabla f(x)$

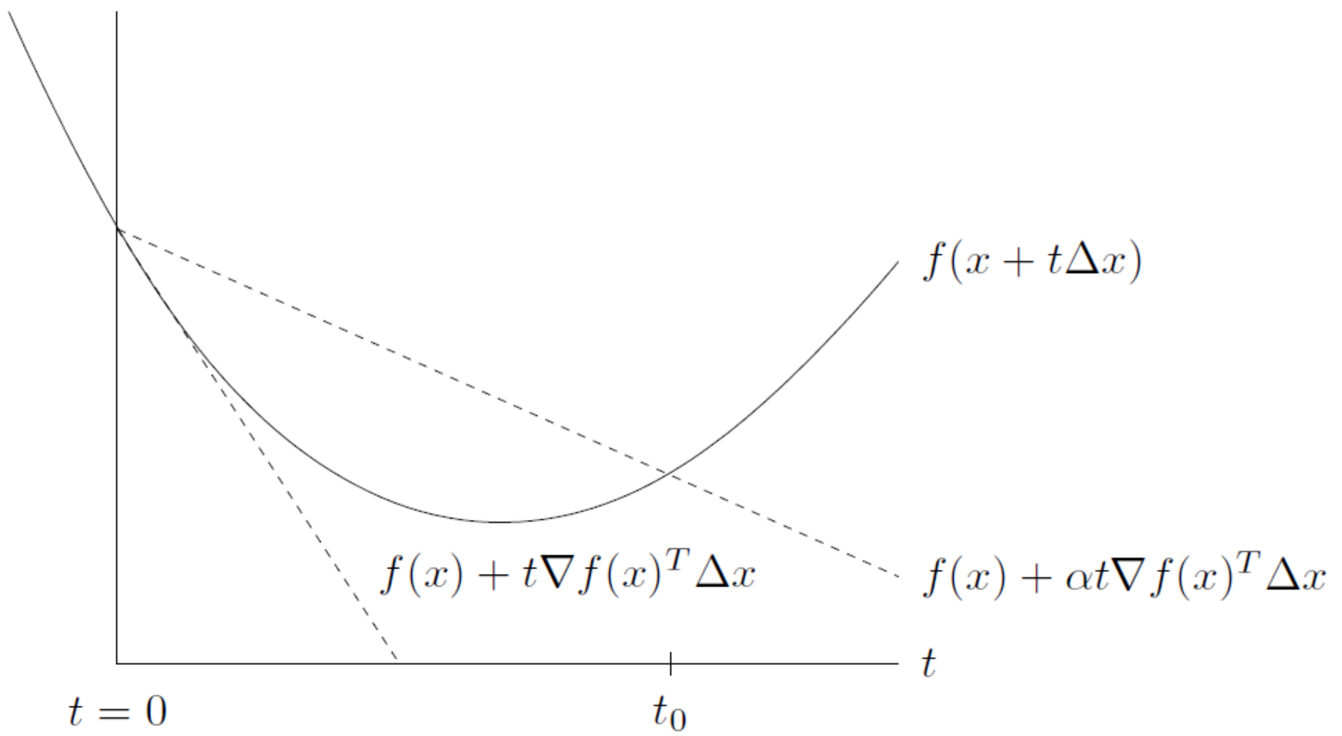
$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

$$\begin{aligned} f(x^+) &= f\big(x - t\nabla f(x)\big) \\ &\geq f(x) + \nabla f(x)^T (x - t\nabla f(x) - x) \\ &= f(x) - t\|\nabla f(x)\|_2^2. \end{aligned}$$

- Lets assume current state is point x, and a step direction $\Delta x = -\nabla f(x)$.
- We would like to find x^+ such that
  $$f(x) \geq f(x^+).$$
- By convexity, the tangent line
  $$f(x) + t\nabla f(x)^T \Delta x$$
  is always lower than $f(x)$.
- Thus, before making a comparison we adjust this value by fraction α, and then compare progress with $f(x) + \alpha t\nabla f(x)^T \Delta x$.

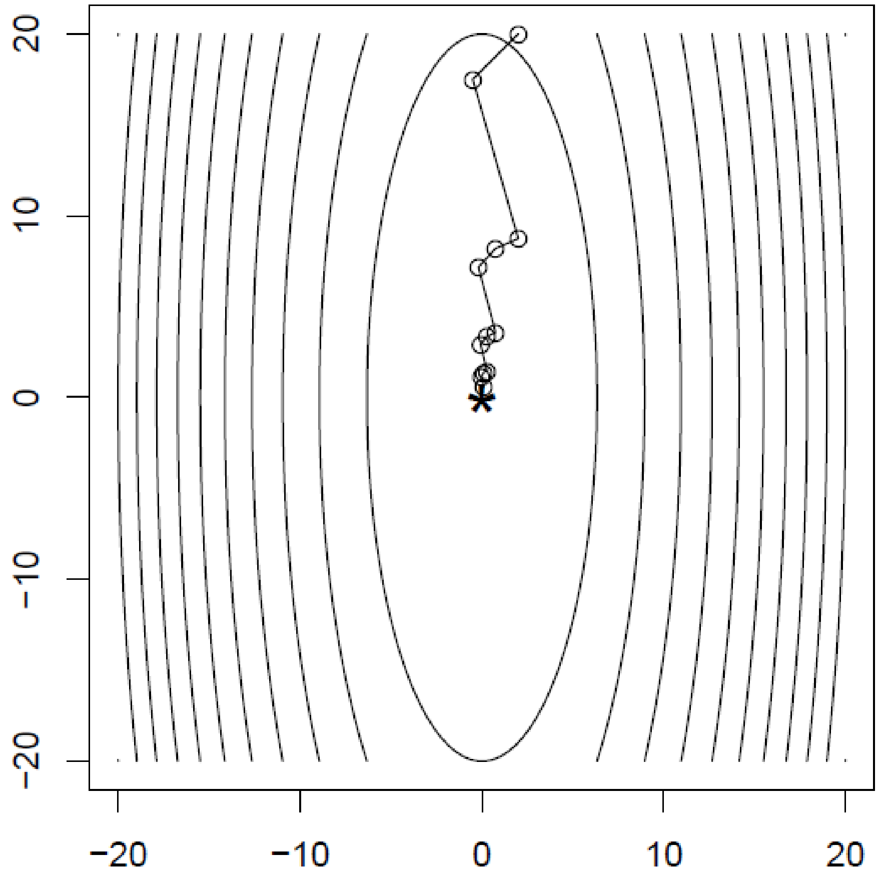
Backtracking interpretation

- If the value of the function in the proposed step $f(x - t\nabla f(x))$ is to big, we adjust by a factor β and repeat until we find a value of f(x^+) that is lower or equal than our benchmark.
- If the criterion is meet, we update our next value to $x^+ = x - t\nabla f(x)$.



For $\Delta x = -\nabla f(x)$

Backtracking picks up roughly the right step size (12 outer steps, 40 steps total):



Here $\alpha = \beta = 0.5$

Convergence analysis

Assume that f convex and differentiable, with $\mathrm{dom}(f) = \mathbb{R}^n$, and additionally

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \text{for any } x, y$$

I.e., ∇f is Lipschitz continuous with constant $L > 0$

Theorem: Gradient descent with fixed step size $t \leq 1/L$ satisfies

$$f(x^{(k)}) - f^\star \leq \frac{\|x^{(0)} - x^\star\|_2^2}{2tk}$$

We say gradient descent has convergence rate $O(1/k)$

Convergence analysis

- *The gradient descent with fixed step size* $t < 1/L$ *satisfies*

$$f(x^{(k)}) - f^* \leq \frac{||x^{(0)} - x^*||_2^2}{2tk}$$

- From this we can see that

$$\epsilon = \frac{||x^{(0)} - x^*||_2^2}{2tk} \implies k = \frac{||x^{(0)} - x^*||_2^2}{2t\epsilon}$$

Hence, $O(1/\epsilon)$ iterations are required for $f(x^{(k)}) - f^* \leq \epsilon$.

Proof: By assumption ∇f is Lipschitz with constant L which implies

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}||y-x||_2^2 \quad \forall\, x, y \qquad (1.1)$$

so we can upper bound the function by a quadratic .

- Suppose we are at a x in gradient descent iterations, go to $x^+ = x - t\nabla f(x)$.

  Evaluating the inequality in 1.1 at $y = x^+$ we find that

$$\begin{aligned}
f(x^+) &\leq f(x) + \nabla f(x)^T(x^+ - x) + \frac{L}{2}||x^+ - x||_2^2 \\
&= f(x) + \nabla f(x)^T(x - t\nabla f(x) - x) + \frac{L}{2}||x - t\nabla f(x) - x||_2^2 \\
&= f(x) - t\nabla f(x)^T(\nabla f(x)) + \frac{L}{2}||t\nabla f(x)||_2^2 \\
&= f(x) - t||\nabla f(x)||_2^2 + \frac{Lt^2}{2}||\nabla f(x)||_2^2 \\
&= f(x) - t(1 - \frac{Lt}{2})||\nabla f(x)||_2^2 \\
&\leq f(x) - \frac{t}{2}||\nabla f(x)||_2^2
\end{aligned}$$

because $t < 1/L$ and hence, $Lt/2 < 1/2$.

- Thus, we have shown that

$$f(x^+) \leq f(x) - \frac{t}{2}||\nabla f(x)||_2^2 \qquad (1.2)$$

or that $f(x^+) < f(x)$ showing descent.

- Since f is convex the first order characterization holds and hence

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) \quad \forall\, x, y \in \text{dom}(f)$$

- Rearranging and setting $y = x^*$ yields

$$f(x) \leq f(x^*) + \nabla f(x)^T(x - x^*) \qquad (1.3)$$

- Combining this with 1.2 we have

$$\begin{aligned} f(x^+) &\leq f(x) - \frac{t}{2}||\nabla f(x)||_2^2 \\ &\leq f(x^*) + \nabla f(x)^T(x - x^*) - \frac{t}{2}||\nabla f(x)||_2^2 \end{aligned}$$

$$
\begin{aligned}
f(x^+) &\le f(x) - \frac{t}{2}||\nabla f(x)||_2^2 \\
&\le f(x^*) + \nabla f(x)^T(x - x^*) - \frac{t}{2}||\nabla f(x)||_2^2 \\
&= f(x^*) + \frac{1}{2t}\Big(||x - x^*||_2^2 - ||x - x^*||_2^2 - t^2||\nabla f(x)||_2^2 + 2t\nabla f(x)^T(x - x^*)\Big) \\
&= f(x^*) + \frac{1}{2t}\Big(||x - x^*||_2^2 - (x - x^*)^T(x - x^*) - t^2\nabla f(x)^T\nabla f(x) + 2t\nabla f(x)^T(x - x^*)\Big) \\
&= f(x^*) + \frac{1}{2t}\Big(||x - x^*||_2^2 - [(x - x^*)^T(x - x^*) + t^2\nabla f(x)^T\nabla f(x) - 2t\nabla f(x)^T(x - x^*)]\Big) \\
&= f(x^*) + \frac{1}{2t}\Big(||x - x^*||_2^2 - [(x - t\nabla f(x)^T - x^*)^T(x - t\nabla f(x)^T - x^*)]\Big) \\
&= f(x^*) + \frac{1}{2t}\Big(||x - x^*||_2^2 - ||x - t\nabla f(x)^T - x^*||_2^2\Big) \\
&= f(x^*) + \frac{1}{2t}\Big(||x - x^*||_2^2 - ||x^+ - x^*||_2^2\Big)
\end{aligned}
$$

because $x^+ = x - t\nabla f(x)$.

$$\boxed{f(x^+) \le f(x^*) + \frac{1}{2t}\left(||x - x^*||_2^2 - ||x^+ - x^*||_2^2\right)}$$

Applying this result to a step i we find that

$$f(x^{(i)}) - f(x^*) \le \frac{1}{2t}\left(||x^{(i-1)} - x^*||_2^2 - ||x^{(i)} - x^*||_2^2\right)$$

Thus,

$$\begin{aligned}
\sum_{i=1}^{k} f(x^{(i)}) - f(x^*) &\le \sum_{i=1}^{k} \frac{1}{2t}\left(||x^{(i-1)} - x^*||_2^2 - ||x^{(i)} - x^*||_2^2\right) \\
&= \frac{1}{2t}\left(||x^{(0)} - x^*||_2^2 - ||x^{(k)} - x^*||_2^2\right) \\
&\le \frac{1}{2t}\left(||x^{(0)} - x^*||_2^2\right)
\end{aligned}$$

The last step follows because this is a telescoping sum where the second term for each $i-1$ cancels with the first term for each i.

recall: $f(x^+) < f(x)$ shows descent.

$$\frac{1}{k}\sum_{i=1}^{k} f(x^{(i)}) - f(x^*) \geq \frac{1}{k}\sum_{i=1}^{k} f(x^{(k)}) - f(x^*) = f(x^{(k)}) - f(x^*)$$

Combining these yields our desired result

$$\sum_{i=1}^{k} f(x^{(i)}) - f(x^*) \geq k\left(f(x^{(k)}) - f(x^*)\right)$$

From previous slide:

$$\boxed{\sum_{i=1}^{k} f(x^{(i)}) - f(x^*) \leq \frac{1}{2t}\left(||x^{(0)} - x^*||_2^2\right)}$$

$$k\left(f(x^{(k)}) - f(x^*)\right) \leq \frac{1}{2t}\left(||x^{(0)} - x^*||_2^2\right)$$

$$\rightarrow \boxed{f(x^{(k)}) - f(x^*) \leq \frac{||x^{(0)} - x^*||_2^2}{2tk}}$$

Stochastic gradient descent

Consider minimizing a sum of functions

$$\min_x \sum_{i=1}^m f_i(x)$$

As $\nabla \sum_{i=1}^m f_i(x) = \sum_{i=1}^m \nabla f_i(x)$, gradient descent would repeat:

$$x^{(k)} = x^{(k-1)} - t_k \cdot \sum_{i=1}^m \nabla f_i(x^{(k-1)}), \quad k = 1, 2, 3, \dots$$

In comparison, stochastic gradient descent or SGD (or incremental gradient descent) repeats:

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f_{i_k}(x^{(k-1)}), \quad k = 1, 2, 3, \dots$$

where $i_k \in \{1, \dots m\}$ is some chosen index at iteration k

Two rules for choosing index i_k at iteration k:

- Cyclic rule: choose $i_k = 1, 2, \ldots m, 1, 2, \ldots m, \ldots$
- Randomized rule: choose $i_k \in \{1, \ldots m\}$ uniformly at random

Randomized rule is more common in practice

What's the difference between stochastic and usual (called batch) methods? Computationally, m stochastic steps $\approx$ one batch step. But what about progress?

- Cyclic rule, m steps: $x^{(k+m)} = x^{(k)} - t \sum_{i=1}^m \nabla f_i(x^{(k+i-1)})$

m stochastic steps

$$x^{(k+1)} = x^{(k)} - t \nabla f_1(x^{(k)})$$

$$x^{(k+2)} = x^{(k+1)} - t \nabla f_2(x^{(k+1)}) = x^{(k)} - t \nabla f_1(x^{(k)}) - t \nabla f_2(x^{(k+1)})$$

$$\vdots$$

$$x^{(k+m)} = x^{(k+m-1)} - t \nabla f_m(x^{(k+m-1)}) = x^{(k)} - t \sum_{i=1}^m \nabla f_i(x^{(k+i-1)})$$

- Batch method, one step: $x^{(k+1)} = x^{(k)} - t\sum_{i=1}^{m} \nabla f_i(x^{(k)})$
- Difference in direction is $\sum_{i=1}^{m}[\nabla f_i(x^{(k+i-1)}) - \nabla f_i(x^{(k)})]$

So SGD should converge if each $\nabla f_i(x)$ doesn't vary wildly with x

Rule of thumb: SGD thrives far from optimum, struggles close to optimum ...

Appendix

Some notes from multi-variate calculus

<u>Lipschitz continuity</u>

- A **Lipschitz continuous function** is limited in how fast it can change:
  - there exists a definite real number such that,
    - for every pair of points on the graph of this function,
      - ✓ the absolute value of the slope of the line connecting them is not greater than this real number.
    - this bound is called a **Lipschitz constant** of the function.
  - For instance, <u>every function that has bounded first derivatives is Lipschitz</u>.

In particular, a real-valued function $f : R \to R$ is called Lipschitz continuous if there exists a positive real constant K such that, for all real x_1 and x_2,

$$|f(x_1) - f(x_2)| \leq K|x_1 - x_2|.$$

the sine function is Lipschitz continuous because its derivative, the cosine function, is bounded above by 1 in absolute value.

Lipschitz continuous gradient

the gradient of f is *Lipschitz continuous* with parameter $L > 0$ if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \text{for all } x, y \in \mathbf{dom}\, f$$

- note that the definition does not assume convexity of f
- we will see that for convex f with $\mathbf{dom}\, f = \mathbf{R}^n$, this is equivalent to

$$\frac{L}{2}x^T x - f(x) \quad \text{is convex}$$

(*i.e.*, if f is twice differentiable, $\nabla^2 f(x) \preceq LI$ for all x)

Cauchy–Schwarz inequality

- The Cauchy–Schwarz inequality states that for all vectors u and v

$$|\langle \mathbf{u}, \mathbf{v} \rangle|^2 \leq \langle \mathbf{u}, \mathbf{u} \rangle \cdot \langle \mathbf{v}, \mathbf{v} \rangle,$$

where $\langle \cdot, \cdot \rangle$ is the inner product.

- Equivalently, by taking the square root of both sides, and referring to the norms of the vectors, the inequality is written as

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \|\mathbf{v}\|.$$

Monotonicity of gradient

a differentiable function f is convex if and only if $\mathbf{dom}\, f$ is convex and

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq 0 \quad \text{for all } x, y \in \mathbf{dom}\, f$$

i.e., the gradient $\nabla f : \mathbf{R}^n \to \mathbf{R}^n$ is a *monotone* mapping

Proof

- if f is differentiable and convex, then

$$f(y) \geq f(x) + \nabla f(x)^T (y - x), \qquad f(x) \geq f(y) + \nabla f(y)^T (x - y)$$

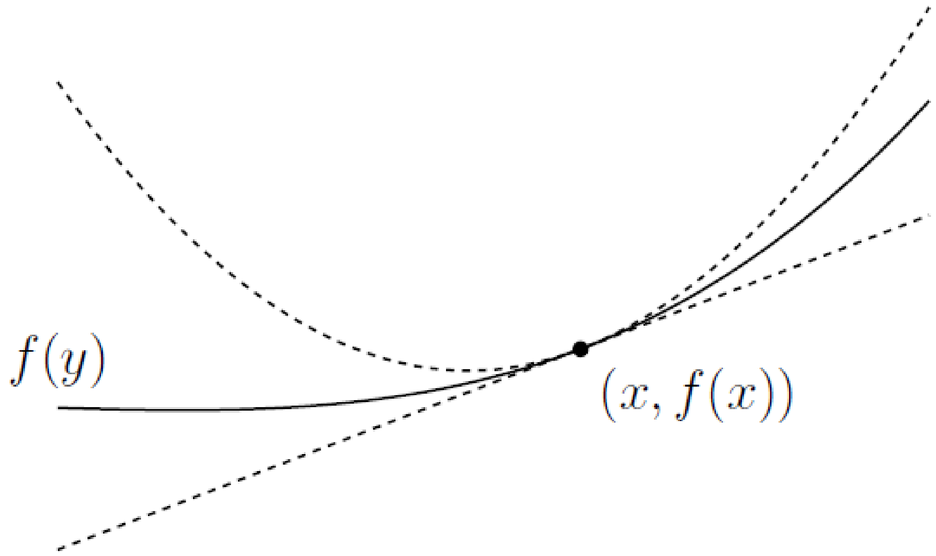
combining the inequalities gives $(\nabla f(x) - \nabla f(y))^T (x - y) \geq 0$

Quadratic upper bound

suppose ∇f is Lipschitz continuous with parameter L and $\mathrm{dom}\, f$ is convex

- then $g(x) = (L/2)x^T x - f(x)$, with $\mathrm{dom}\, g = \mathrm{dom}\, f$, is convex
- convexity of g is equivalent to a quadratic upper bound on f:

$$f(y) \le f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|_2^2 \quad \text{for all } x, y \in \mathrm{dom}\, f$$



Proof of Quadratic Upper Bound

- f is convex $\rightarrow$ $(\nabla f(x) - \nabla f(y))^T (x - y) \geq 0$

- the Cauchy-Schwarz inequality imply

$$(\nabla f(x) - \nabla f(y))^T (x - y) \leq \|(\nabla f(x) - \nabla f(y))^T\|_2 \|x - y\|_2$$

- Lipschitz continuity of ∇f $\rightarrow$ $\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$

$$\boxed{(\nabla f(x) - \nabla f(y))^T (x - y) \leq L\|x - y\|_2^2 \quad \text{for all } x, y \in \operatorname{dom} f}$$

Proof of Quadratic Upper Bound

$$(\nabla f(x) - \nabla f(y))^T (x - y) \le L\|x - y\|_2^2 \quad \text{for all } x, y \in \operatorname{dom} f$$

To prove a quadratic upper bound for f(x), we first prove g(x) is convex.

$$g(x) = (L/2)x^T x - f(x),$$

$$\nabla g(x) = Lx - \nabla f(x)$$

$$(\nabla g(x) - \nabla g(y))^T (x - y) =$$

$$\left[L\,(x - y)^T - (\nabla f(x) - \nabla f(y))^T \right] (x - y) =$$

$$L\|x - y\|_2^2 \; - \; (\nabla f(x) - \nabla f(y))^T \, (x - y) \; \ge 0$$

Hence, g(x) is convex!

Proof of Quadratic Upper Bound

- the quadratic upper bound is the first-order condition for convexity of g

$$g(y) \geq g(x) + \nabla g(x)^T(y - x) \quad \text{for all } x, y \in \operatorname{dom} g$$

Replace the following values in the above expression:

$$g(x) = (L/2)x^T x - f(x),$$
$$g(y) = (L/2)y^T y - f(y)$$

$$\nabla g(x) = Lx - \nabla f(x)$$

$$\nabla g(y) = Ly - \nabla f(y)$$

You'll obtain the quadratic upper bound for f:

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|_2^2 \quad \text{for all } x, y \in \operatorname{dom} f$$