Lecture 08

Subgradients &

the subgradient method

Subgradients

Recall that for convex and differentiable f,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \text{for all } x, y$$

I.e., linear approximation always underestimates f

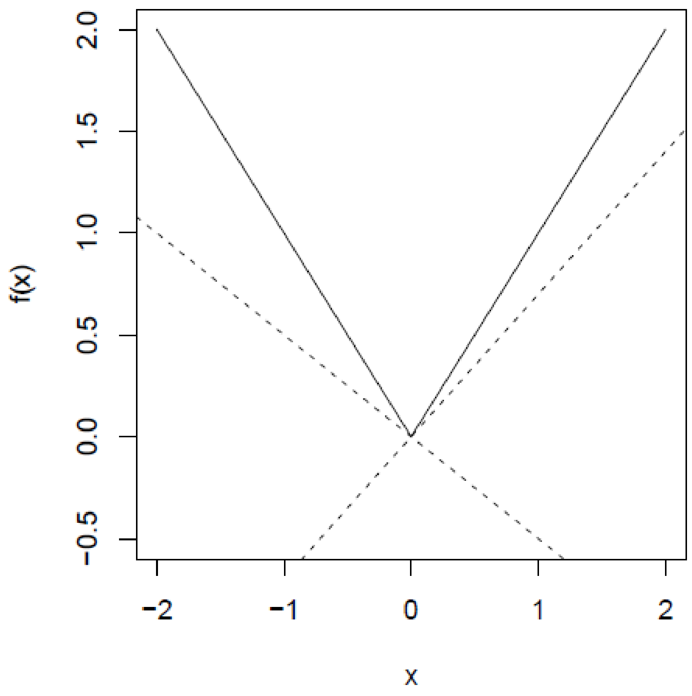
A subgradient of a convex function f at x is any $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + g^T (y - x) \quad \text{for all } y$$

- Always exists
- If f differentiable at x, then $g = \nabla f(x)$ uniquely
- Actually, same definition works for nonconvex f (however, subgradients need not exist)

Examples of subgradients

Consider $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$

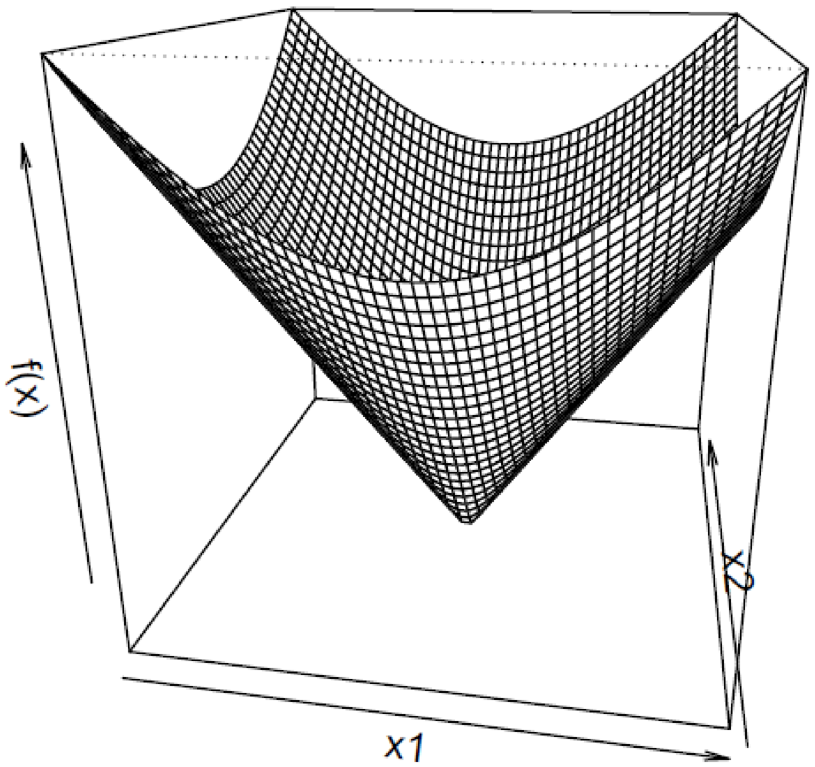


- For $x \neq 0$, unique subgradient $g = \text{sign}(x)$
- For $x = 0$,

$$|y| \geq |x| + g^T(y - x)$$

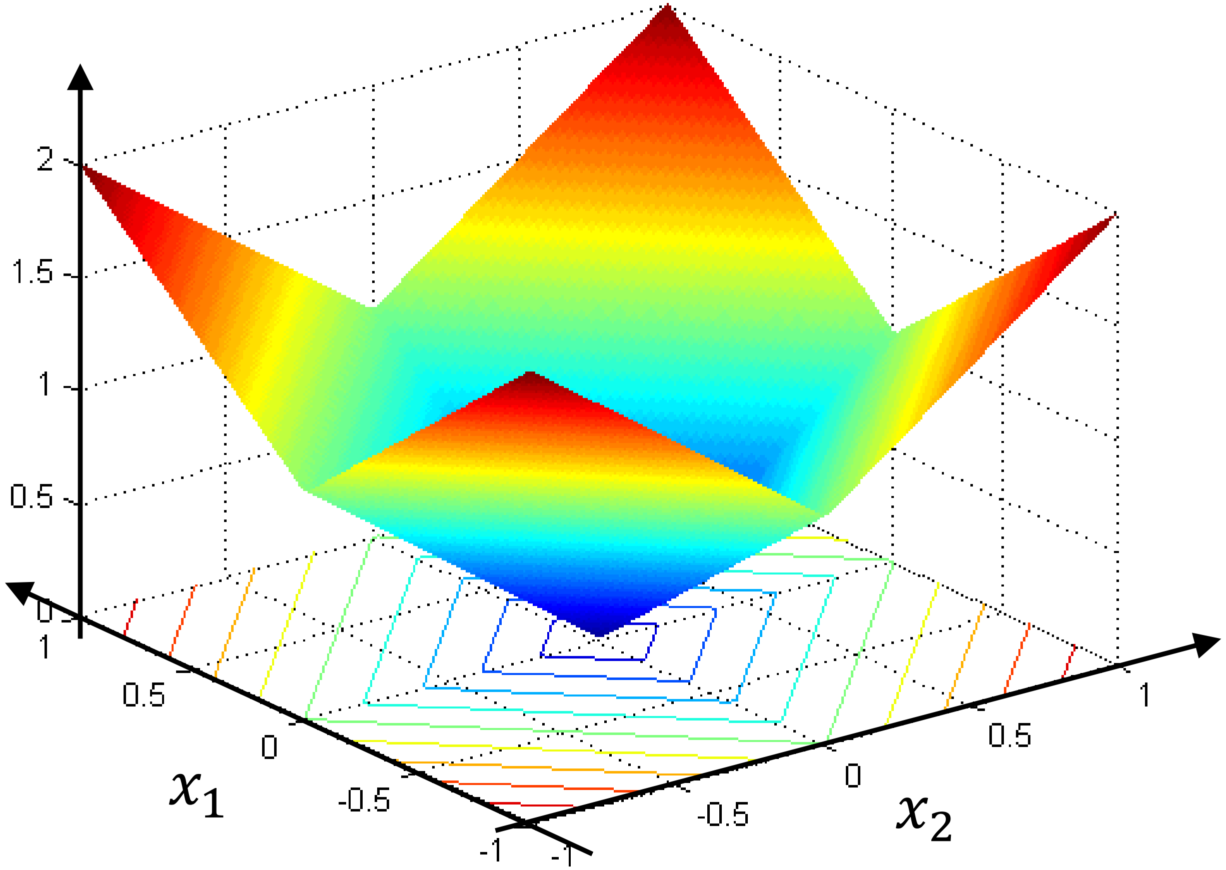
$$|y| \geq g^T y \implies g \in [-1, 1].$$

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_2$



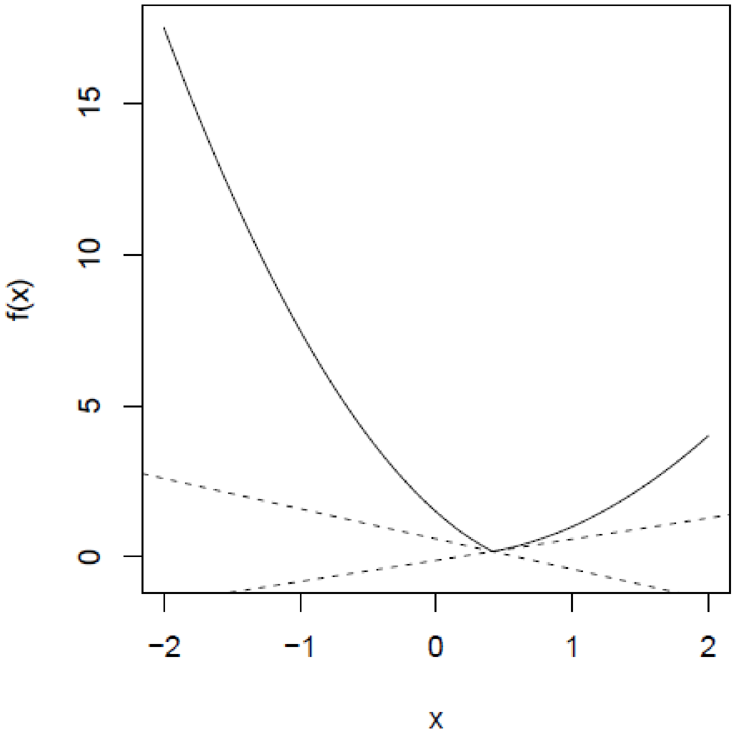
- For $x \neq 0$, unique subgradient $g = x/\|x\|_2$
- For $x = 0$, subgradient g is any element of $\{z : \|z\|_2 \leq 1\}$

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_1$



- For $x_i \neq 0$, unique ith component $g_i = \operatorname{sign}(x_i)$
- For $x_i = 0$, ith component g_i is any element of $[-1, 1]$

Let $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable, and consider $f(x) = \max\{f_1(x), f_2(x)\}$



- For $f_1(x) > f_2(x)$, unique subgradient $g = \nabla f_1(x)$
- For $f_2(x) > f_1(x)$, unique subgradient $g = \nabla f_2(x)$
- For $f_1(x) = f_2(x)$, subgradient g is any point on the line segment between $\nabla f_1(x)$ and $\nabla f_2(x)$

Subdifferential

Set of all subgradients of convex f is called the subdifferential:

$$\partial f(x) = \{g \in \mathbb{R}^n : g \text{ is a subgradient of } f \text{ at } x\}$$

- $\partial f(x)$ is closed and convex (even for nonconvex f)
- Nonempty (can be empty for nonconvex f)
- If f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$
- If $\partial f(x) = \{g\}$, then f is differentiable at x and $\nabla f(x) = g$

- Suppose f is the pointwise maximum of convex functions $f_1, \dots, f_m$, *i.e.*,

$$f(x) = \max_{i=1,\dots,m} f_i(x),$$

where the functions f_i are subdifferentiable.

- We first show how to construct a subgradient of f at x.
- Let k be any index for which $f_k(x) = f(x)$, and let $g \in \partial f_k(x)$.

Then $g \in \partial f(x)$.

- In other words, to find a subgradient of the maximum of functions, we can choose one of the functions that achieves the maximum at the point, and choose any subgradient of that function at the point.

This follows from

$$f(z) \geq f_k(z) \geq f_k(x) + g^T(z - x) = f(x) + g^T(y - x).$$

- More generally, we have

$$\partial f(x) = \mathbf{Co} \cup \{\partial f_i(x) \mid f_i(x) = f(x)\},$$

i.e., the subdifferential of the maximum of functions is the convex hull of the union of subdifferentials of the 'active' functions at x.

Subgradient calculus

Basic rules for convex functions:

- Scaling: $\partial(af) = a \cdot \partial f$ provided $a > 0$
- Addition: $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$
- Affine composition: if $g(x) = f(Ax + b)$, then

$$\partial g(x) = A^T \partial f(Ax + b)$$

- Finite pointwise maximum: if $f(x) = \max_{i=1,\ldots m} f_i(x)$, then

$$\partial f(x) = \operatorname{conv}\left(\bigcup_{i: f_i(x) = f(x)} \partial f_i(x) \right)$$

convex hull of union of subdifferentials of all active functions at x

Optimality condition

For any f (convex or not),

$$f(x^\star) = \min_x \; f(x) \quad \Longleftrightarrow \quad 0 \in \partial f(x^\star)$$

I.e., $x^\star$ is a minimizer if and only if 0 is a subgradient of f at $x^\star$. This is called the subgradient optimality condition

Why? Easy: $g = 0$ being a subgradient means that for all y

$$f(y) \geq f(x^\star) + 0^T(y - x^\star) = f(x^\star)$$

Note the implication for a convex and differentiable function f, with $\partial f(x) = \{\nabla f(x)\}$

Connection to convex geometry

Convex set $C \subseteq \mathbb{R}^n$, consider indicator function $I_C : \mathbb{R}^n \to \mathbb{R}$,

$$I_C(x) = I\{x \in C\} = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C \end{cases}$$

For $x \in C$, $\partial I_C(x) = \mathcal{N}_C(x)$, the normal cone of C at x, recall

$$\mathcal{N}_C(x) = \{g \in \mathbb{R}^n : g^T x \geq g^T y \text{ for any } y \in C\}$$

Why? By definition of subgradient g,

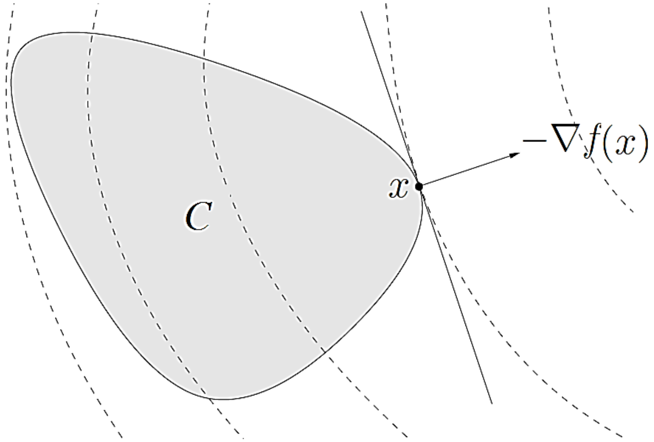
$$I_C(y) \geq I_C(x) + g^T (y - x) \quad \text{for all } y$$

- For $y \notin C$, $I_C(y) = \infty$
- For $y \in C$, this means $0 \geq g^T (y - x)$

Derivation of first-order optimality

Example of the power of subgradients: we can use what we have learned so far to derive the first-order optimality condition. Recall that for f convex and differentiable, the problem

$$\min_x \; f(x) \;\text{ subject to }\; x \in C$$



is solved at x if and only if

$$\nabla f(x)^T (y - x) \geq 0 \;\text{ for all }\; y \in C$$

Intuitively says that gradient increases as we move away from x. How to see this? First recast problem as

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

Now apply subgradient optimality: $0 \in \partial(f(x) + I_C(x))$

But

$$
\begin{aligned}
0 \in \partial\big(f(x) + I_C(x)\big) \\
&\iff 0 \in \{\nabla f(x)\} + \mathcal{N}_C(x) \\
&\iff -\nabla f(x) \in \mathcal{N}_C(x) \\
&\iff -\nabla f(x)^T x \geq -\nabla f(x)^T y \text{ for all } y \in C \\
&\iff \nabla f(x)^T (y - x) \geq 0 \text{ for all } y \in C
\end{aligned}
$$

as desired

Note: the condition $0 \in \partial f(x) + \mathcal{N}_C(x)$ is a fully general condition for optimality in a convex problem. But this is not always easy to work with (KKT conditions, later, are easier)

Example: optimality conditions

Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$,

$$\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

where $\lambda \geq 0$. Subgradient optimality:

$$\begin{aligned} 0 \in \partial\Big(\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\Big) & \\ \iff & \quad 0 \in \Big(-X^T(y - X\beta) + \lambda\partial\|\beta\|_1\Big) \\ \iff & \quad X^T(y - X\beta) = \lambda v \end{aligned}$$

for some $v \in \partial\|\beta\|_1$, i.e.,

$$v_i \in \begin{cases} \{1\} & \text{if } \beta_i > 0 \\ \{-1\} & \text{if } \beta_i < 0 \, , \quad i = 1, \ldots p \\ [-1, 1] & \text{if } \beta_i = 0 \end{cases}$$

Write $X_1, \ldots X_p$ for columns of X. Then subgradient optimality reads:

$$\begin{cases} X_i^T(y - X\beta) = \lambda \cdot \mathrm{sign}(\beta_i) & \text{if } \beta_i \neq 0 \\ |X_i^T(y - X\beta)| \leq \lambda & \text{if } \beta_i = 0 \end{cases}$$

Note: the subgradient optimality conditions do not directly lead to an expression for a solution ... however they do provide a way to check optimality

Example: distance to a convex set

Recall the distance function to a closed, convex set C:

$$\operatorname{dist}(x, C) = \min_{y \in C} \|y - x\|_2$$

This is a convex function. What are its subgradients?

Write $\operatorname{dist}(x, C) = \|x - P_C(x)\|_2$, where $P_C(x)$ is the projection of x onto C. It turns out that when $\operatorname{dist}(x, C) > 0$,

$$\partial \operatorname{dist}(x, C) = \left\{ \frac{x - P_C(x)}{\|x - P_C(x)\|_2} \right\}$$

Only has one element, so in fact $\operatorname{dist}(x, C)$ is differentiable and this is its gradient

We will only show one direction, i.e., that

$$\frac{x - P_C(x)}{\|x - P_C(x)\|_2} \in \partial \mathrm{dist}(x, C)$$

Write $u = P_C(x)$. Then by first-order optimality conditions for a projection,

$$(x-u)^T(y-u) \le 0 \quad \text{for all } y \in C$$

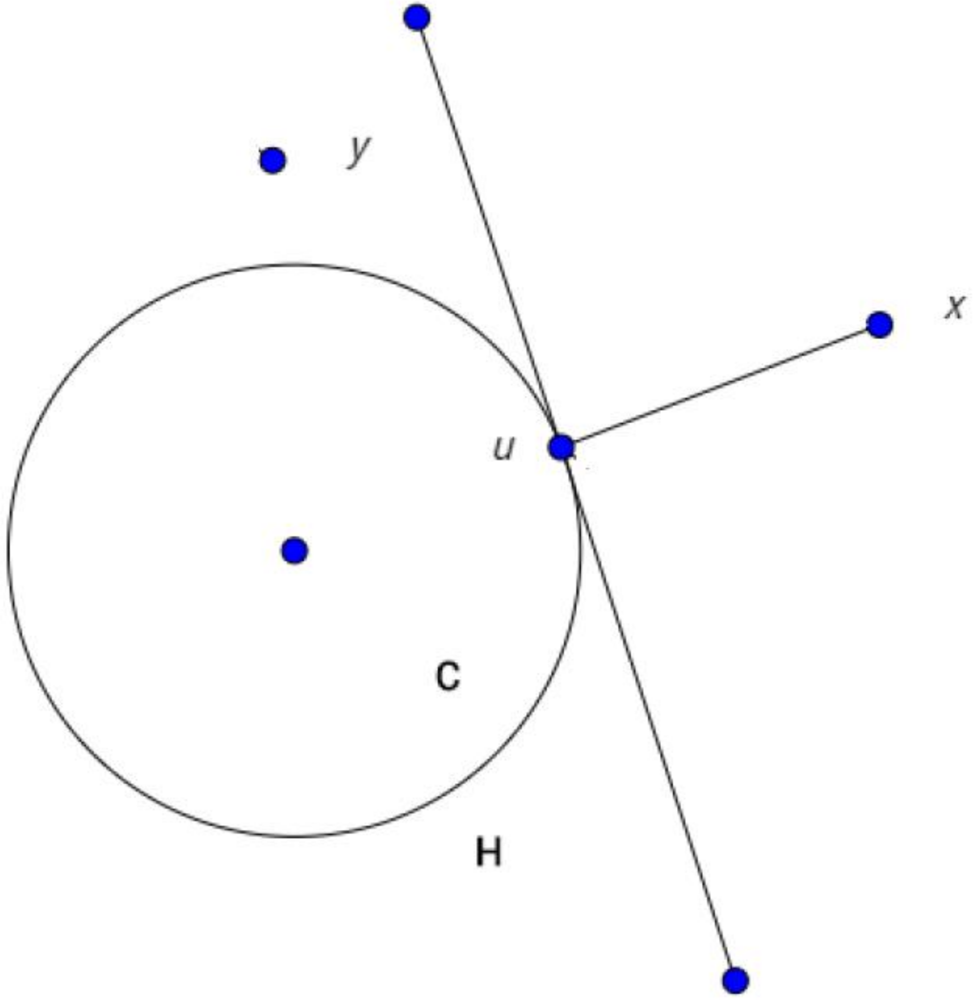
Hence

$$C \subseteq H = \{y : (x-u)^T(y-u) \le 0\}$$

Claim:

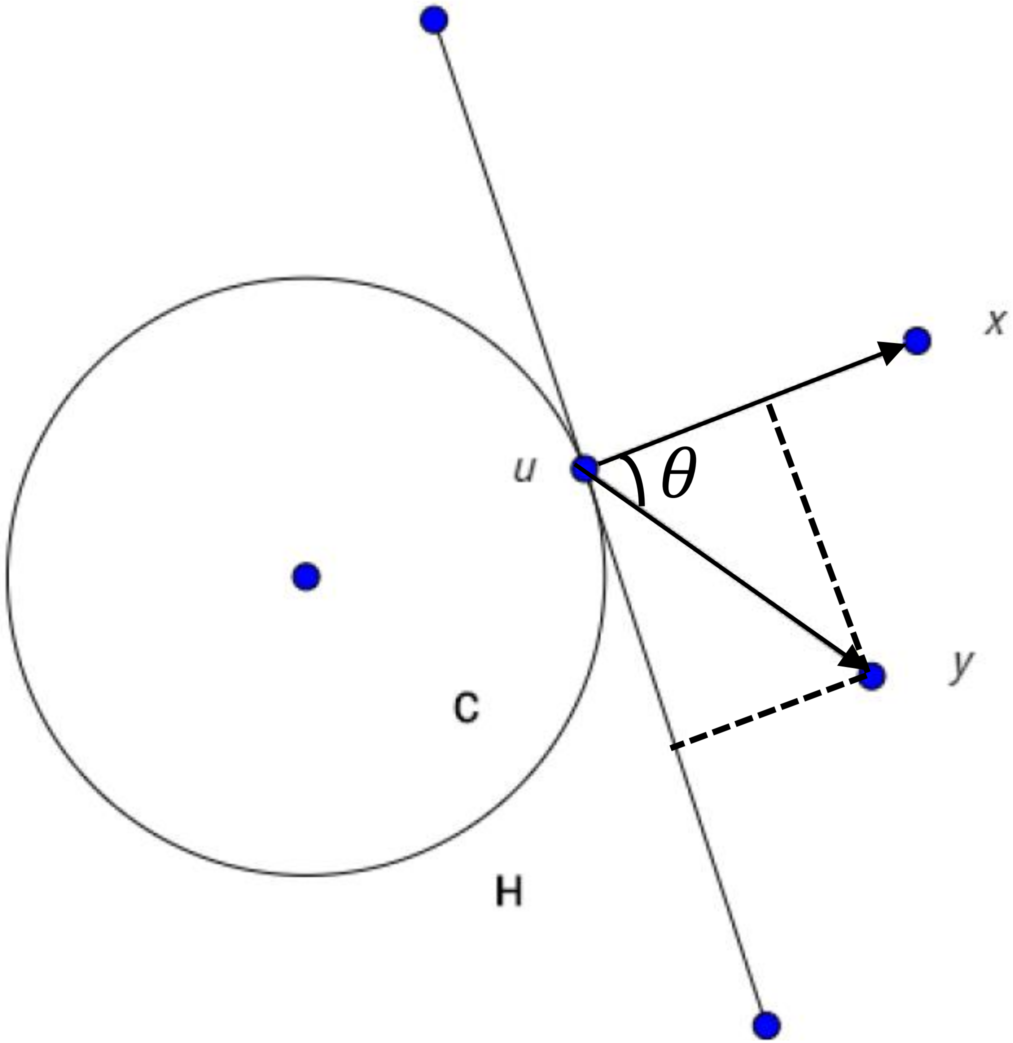
$$\mathrm{dist}(y, C) \ge \frac{(x-u)^T(y-u)}{\|x-u\|_2} \quad \text{for all } y$$

Check: first, for $y \in H$, the right-hand side is ≤ 0



For $y \in H$,

$$(x-u)^T(y-u) \leq 0$$



For $y \notin H$,

$$(x-u)^T(y-u) = ||x-u||_2||y-u||_2 \cos\theta,$$

where θ is the angle between $x-u$ and $y-u$.

(i) For $y \in H$,

$$(x-u)^T(y-u) \leq 0$$

$$\text{dist}(y,C) \geq 0$$

$$\text{dist}(y,C) \geq \frac{(x-u)^T(y-u)}{||x-u||_2}$$

(ii) For $y \notin H$,

$$\frac{(x-u)^T(y-u)}{||x-u||_2} = ||y-u||_2 \cos\theta = \text{dist}(y,H) \leq \text{dist}(y,C)$$

Therefore, for any y,

$$\text{dist}(y,C) \geq \frac{(x-u)^T(y-u)}{||x-u||_2}$$

for any y,

$$\text{dist}(y, C) \geq \frac{(x-u)^T(y-u)}{||x-u||_2}$$

$$= \frac{(x-u)^T(y-x+x-u)}{||x-u||_2}$$

$$= ||x-u||_2 + (\frac{x-u}{||x-u||_2})^T(y-x)$$

Hence, $g = \frac{x-u}{||x-u||_2}$ is a subgradient of $\text{dist}(x, C)$ at x.

The subgradient method

Recall : gradient descent

Consider the problem

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

for f convex and differentiable, $\mathrm{dom}(f) = \mathbb{R}^n$. Gradient descent: choose initial $x^{(0)} \in \mathbb{R}^n$, repeat:

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

Step sizes t_k chosen to be fixed and small, or by backtracking line search

If ∇f Lipschitz, gradient descent has convergence rate $O(1/\epsilon)$

Downsides:

- Requires f differentiable
- Can be slow to converge

Subgradient method

Now consider f convex, with $\mathrm{dom}(f) = \mathbb{R}^n$, but not necessarily differentiable

Subgradient method: like gradient descent, but replacing gradients with subgradients. I.e., initialize $x^{(0)}$, repeat:

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

where $g^{(k-1)} \in \partial f(x^{(k-1)})$, any subgradient of f at $x^{(k-1)}$

Subgradient method is not necessarily a descent method, so we keep track of best iterate $x^{(k)}_{\text{best}}$ among $x^{(0)}, \ldots x^{(k)}$ so far, i.e.,

$$f(x^{(k)}_{\text{best}}) = \min_{i=0,\ldots k} f(x^{(i)})$$

Step size choices

- Fixed step sizes: $t_k = t$ all $k = 1, 2, 3, \ldots$
- Diminishing step sizes: choose to meet conditions

$$\sum_{k=1}^{\infty} t_k^2 < \infty, \quad \sum_{k=1}^{\infty} t_k = \infty,$$

i.e., square summable but not summable

Important that step sizes go to zero, but not too fast

Other options too, but important difference to gradient descent: step sizes are typically pre-specified, not adaptively computed

Convergence analysis

Assume that f convex, $\mathrm{dom}(f) = \mathbb{R}^n$, and also that f is Lipschitz continuous with constant $G > 0$, i.e.,

$$|f(x) - f(y)| \le G\|x - y\|_2 \quad \text{for all } x, y$$

Theorem: For a fixed step size t, subgradient method satisfies

$$\lim_{k\to\infty} f(x_{\text{best}}^{(k)}) \le f^\star + G^2 t/2$$

Theorem: For diminishing step sizes, subgradient method satisfies

$$\lim_{k\to\infty} f(x_{\text{best}}^{(k)}) = f^\star$$

subgradient method has convergence rate $O(1/\epsilon^2)$ ... compare this to $O(1/\epsilon)$ rate of gradient descent

Example

Given $(x_i, y_i) \in \mathbb{R}^p \times \{0,1\}$ for $i = 1, \dots n$, consider :

$$f(\beta) = \sum_{i=1}^{n} \Big(-y_i x_i^T \beta + \log(1 + \exp(x_i^T \beta) \Big)$$

This is a smooth and convex, with

$$\nabla f(\beta) = \sum_{i=1}^{n} \big(p_i(\beta) - y_i \big) x_i$$

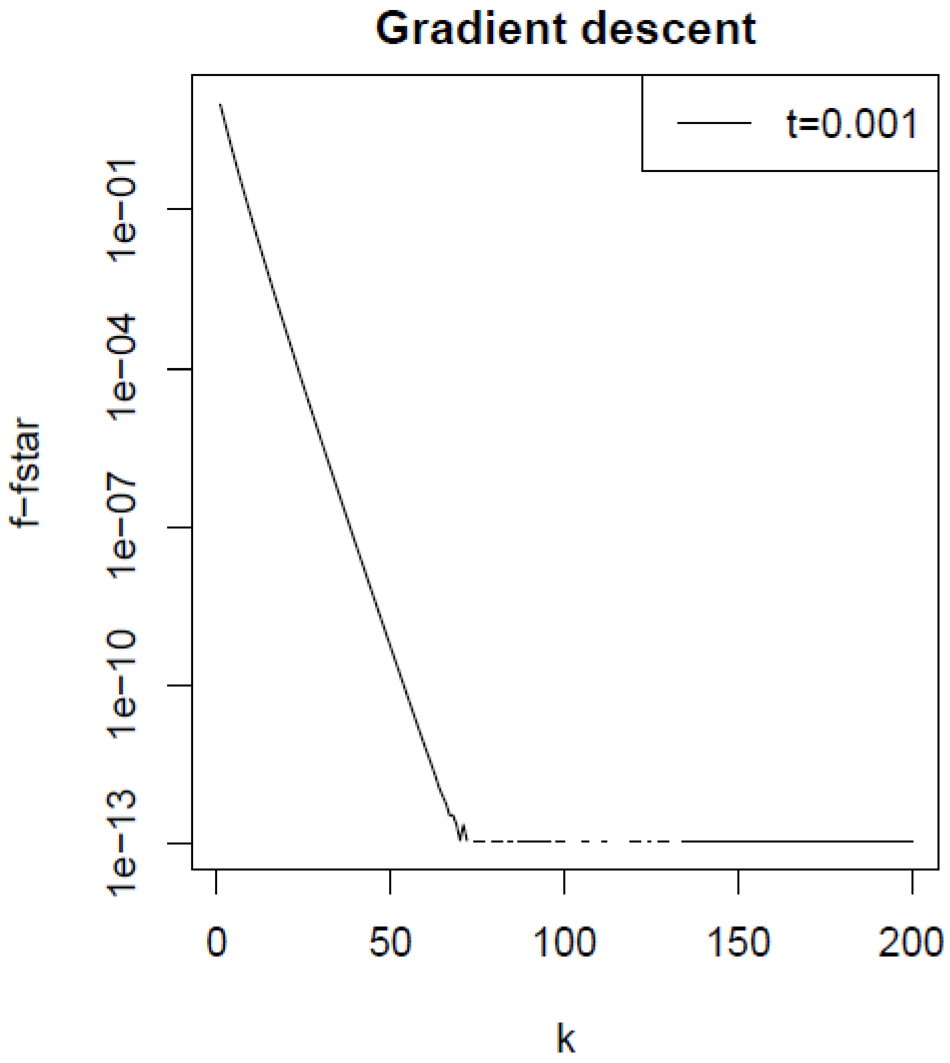
where $p_i(\beta) = \exp(x_i^T \beta)/(1 + \exp(x_i^T \beta))$, $i = 1, \dots n$. We will consider the problem:

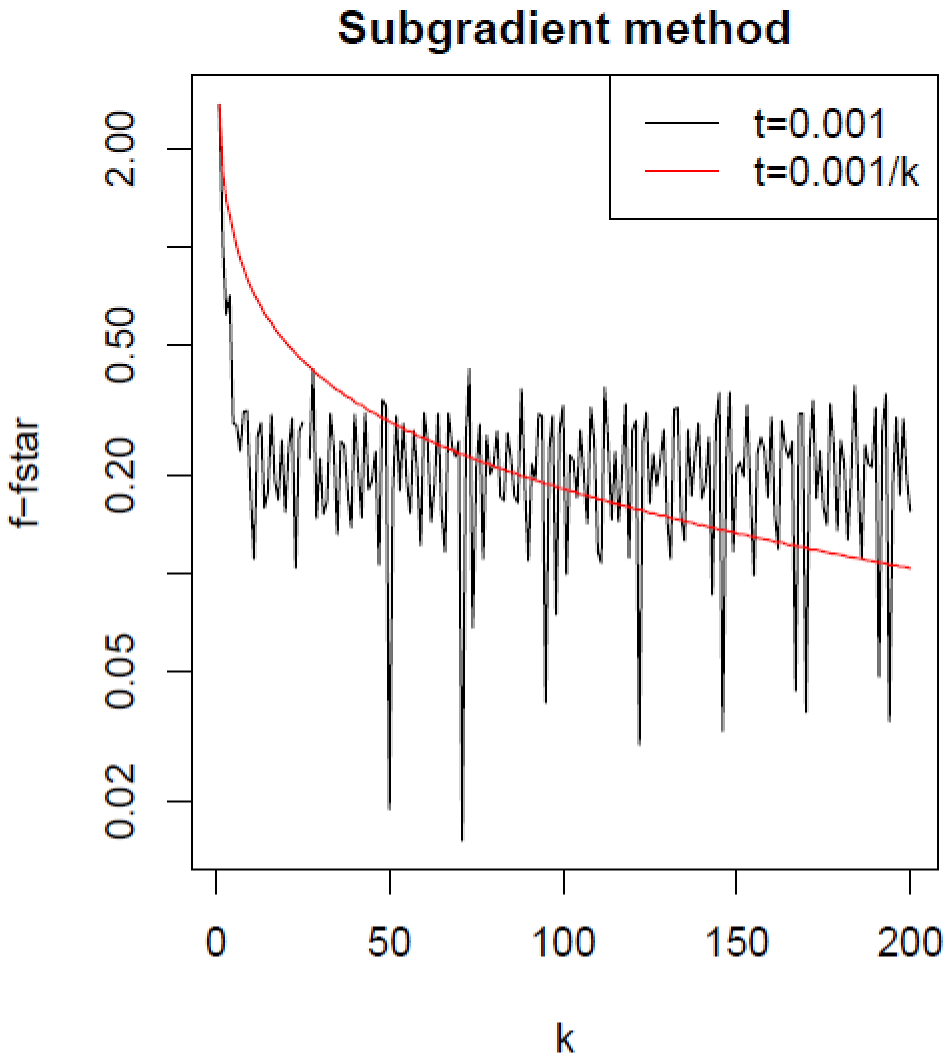
$$\min_{\beta} \; f(\beta) + \lambda \cdot P(\beta)$$

where $P(\beta) \underset{①}{=} \|\beta\|_2^2$ or $P(\beta) \underset{②}{=} \|\beta\|_1$

problem ① : use gradients; problem ② : use subgradients.

Data example with $n = 1000$, $p = 20$:





Stochastic subgradient method

Similar to our setup for stochastic gradient descent. Consider sum of convex functions

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

Stochastic subgradient method repeats:

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

where $i_k \in \{1, \ldots m\}$ is some chosen index at iteration k, chosen by either by the random or cyclic rule, and $g_i^{(k-1)} \in \partial f_i(x^{(k-1)})$ (this update direction is used in place of the usual $\sum_{i=1}^m g_i^{(k-1)}$)

Note that when each f_i, $i = 1, \ldots, m$ is differentiable, this reduces to stochastic gradient descent (SGD)

Convergence of stochastic methods

Assume each f_i, $i = 1, \ldots m$ is convex and Lipschitz with constant $G > 0$

For fixed step sizes $t_k = t$, $k = 1, 2, 3, \ldots$, cyclic and randomized stochastic subgradient methods both satisfy

$$\lim_{k \to \infty} f(x_{\text{best}}^{(k)}) \le f^\star + 5m^2G^2t/2$$

Note: mG can be viewed as Lipschitz constant for whole function $\sum_{i=1}^m f_i$, so this is comparable to batch bound

For diminishing step sizes, cyclic and randomized methods satisfy

$$\lim_{k \to \infty} f(x_{\text{best}}^{(k)}) = f^\star$$

Example

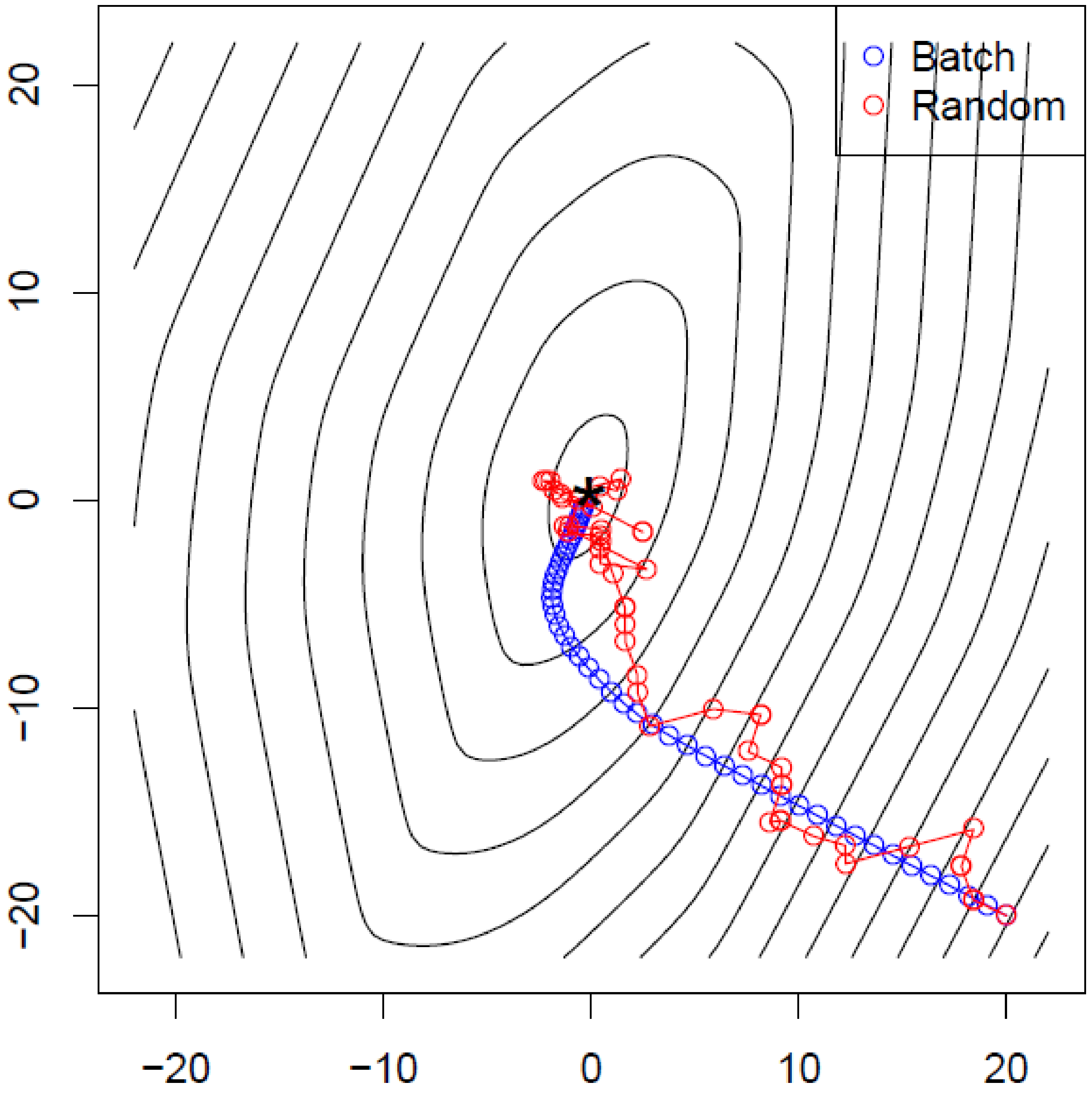
$$\min_{\beta}\ f(\beta) = \sum_{i=1}^{n} \underbrace{\Big(- y_i x_i^T \beta + \log(1 + \exp(x_i^T \beta)\Big)}_{f_i(\beta)}$$

The gradient computation $\nabla f(\beta) = \sum_{i=1}^{n} (\, p_i(\beta) - y_i \,) x_i$ is doable

when n is moderate, but not when $n \approx$ 500 million. Recall:

- One batch update costs $O(np)$
- One stochastic update costs $O(p)$

So clearly, e.g., 10K stochastic steps are much more affordable



<span style="color:blue">Blue</span>: batch steps, $O(np)$
Red: stochastic steps, $O(p)$

Rule of thumb for stochastic methods:

- generally thrive far from optimum
- generally struggle close to optimum

Appendix

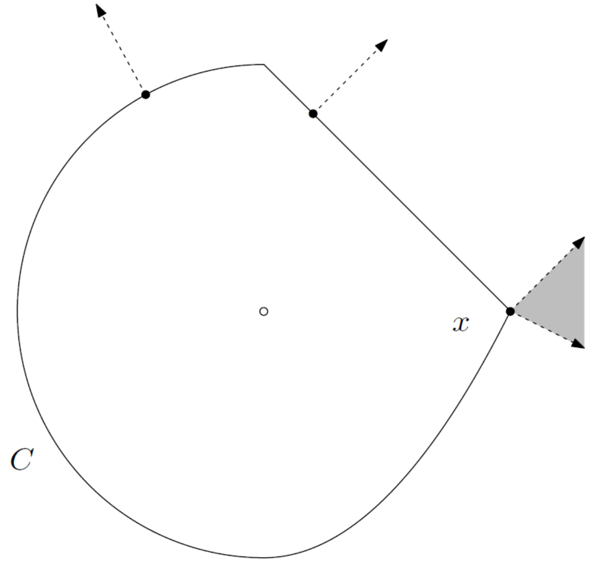
Some notes from convex geometry

“Normal cone” of an arbitrary set

Normal cone: given any set C and point $x \in C$, we can define normal cone as

$$N_C(x) = \{g : g^T x \geq g^T y \text{ for all } y \in C\}$$

- The *normal cone* of a set C at a boundary point x is the set of all vectors g such that $g^T(y - x) \leq 0$ for all $x \in C$ ($i.e.$, the set of vectors that define a supporting hyperplane to C at x).



- Proof: For $g_1, g_2 \in N_C(x)$,

$$(t_1 g_1 + t_2 g_2)^T x = t_1 g_1^T x + t_2 g_2^T x \geq t_1 g_1^T y + t_2 g_2^T y$$
$$= (t_1 g_1 + t_2 g_2)^T y \text{ for all } t_1, t_2 \geq 0$$