Optimization

Basic knowledge

1. *Subgradients*

Definition 1.2 (Subgradients). Let $\mathcal{X} \subset \mathbb{R}^n$, and $f : \mathcal{X} \to \mathbb{R}$. Then $g \in \mathbb{R}^n$ is a subgradient of f at $x \in \mathcal{X}$ if for any $y \in \mathcal{X}$ one has

$$f(x) - f(y) \leq g^\top (x - y).$$

The set of subgradients of f at x is denoted $\partial f(x)$.

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

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

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

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

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

First order optimality

For problem

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

the condition $0 \in \partial f(x) + \mathcal{N}_C(x)$ is a fully general condition for optimality in a convex problem.

Some examples

A. The distance to a convex set

Distance function to a convex set C is

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

then when $\text{dist}(x, C) > 0, \partial \,\text{dist}(x, C) = \left\{ \frac{x - P_C(x)}{\left\| x - P_C(x) \right\|_2} \right\}$.

B. ℓ_0 ***norm***

$f(x) = ||x||_0 = \sum_{i=1}^{n} 1\{x_i \neq 0\}$

ℓ_0 norm function has subgradient at exactly one point: the origin, where $\partial f(0) = 0$. Everywhere else, $\partial f(x) = \emptyset$.

C. ℓ_∞ ***norm***

$f(\boldsymbol{x}) = \|\boldsymbol{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$

If $x \neq 0$***, then pick any*** x_j ***obeying*** $|x_j| = \max_i |x_i|$ ***to get*** $\text{sign}(x_j) e_j \in \partial f(x)$

D. Maximum eigenvalue

$f(\boldsymbol{x}) = \lambda_{max}(x_1 A_1 + \ldots + x_n A_n)$, where $A_1, \ldots, A_n$ are real symmetric matrices.

Rewrite $f(\boldsymbol{x}) = \sup_{\boldsymbol{y}:\|\boldsymbol{y}\|_2 = 1} \boldsymbol{y}^T (x_1 A_1 + \ldots + x_n A_n) \boldsymbol{y}$ as supremum of affine functions of x.

Therefore, taking $\boldsymbol{y}$ as leading eigenvector of $x_1 A_1 + \ldots + x_n A_n$, we have

$$[\boldsymbol{y}^T A_1 \boldsymbol{y}, \ldots, \boldsymbol{y}^T A_n \boldsymbol{y}] \in \partial f(x).$$

2. Oracle complexity

In the black-box model one tries to minimize an unknown function f over a constraint set $\mathcal{X} \subset \mathbb{R}^n$. A black-box optimization procedure is simply a sequence of mappings $\phi_t : \mathcal{X}^t \times \left(\mathbb{R}^n\right)^t \times \mathbb{R}^t \to \mathcal{X}$. The algorithm given by these mappings runs iteratively as follows: initially it makes a query to the point $x_0 = \phi_0(\varnothing)$, and the t^{th} step it queries

$$x_t = \phi_t\left(x_0, \ldots, x_{t-1}, \nabla f\left(x_0\right), \ldots, \nabla f\left(x_{t-1}\right), f\left(x_0\right), \ldots, f\left(x_{t-1}\right)\right).$$

The minimax oracle optimization error after t steps over a set of functions $\mathcal{F}$, is defined as follows

$$\mathrm{OC}_t(\mathcal{F}) = \inf_{\phi_0, \ldots, \phi_t \in \mathcal{F}} \left(f\left(x_t\right) - \inf_{x \in \mathcal{X}} f(x)\right).$$

Intuitively $\mathrm{OC}_t(\mathcal{F})$ describes the best possible ate of convergence for the optimization error when one restricts to black-box procedures that only know the constraint set $\mathcal{X}$ and the class of functions $\mathcal{F}$.

3. *The conjugate function*

Let $f : \mathbb{R}^n \to \mathbb{R}$. The function $f^* : \mathbb{R}^n \to \mathbb{R}$, defined as

$$f^*(\boldsymbol{x}) := \sup_{z \in \mathrm{dom}\, f} \{\langle \boldsymbol{z}, \boldsymbol{x} \rangle - f(\boldsymbol{z})\}$$

is called the conjugate function of f.

Fenchel's inequality

$$x'y \leq f(x) + f^\star(y), \quad \forall x \in \mathfrak{R}^n, y \in \mathfrak{R}^n.$$

The second conjugate

$$f^{**}(x) = \sup_{y \in \mathrm{dom}\, f^*} \left(x^T y - f^*(y)\right).$$

- $f^{**}(x)$ is closed and convex
- From Fenchel's inequality, $x^T y - f^*(y) \leq f(x)$. Hence, $f^{**}(x) \leq f(x)$
- If f is closed and convex, then $f^{**}(x) = f(x)$ for all x; Equivalently, epi f^{**} = epi f.

Conjugate Subgradient Theorem

If f is closed and convex, then the following relations are equivalent for a pair of vectors (x, y):

(i) $x'y = f(x) + f^\star(y)$

(ii) $y \in \partial f(x)$

(iii) $x \in \partial f^\star(y)$

Proof. if $y \in \partial f(x)$, then $f^*(y) = \sup_u (y^T u - f(u)) = y^T x - f(x)$; hence

$$\begin{aligned} f^*(v) &= \sup_u (v^T u - f(u)) \\ &\geq v^T x - f(x) \\ &= x^T(v - y) - f(x) + y^T x \\ &= f^*(y) + x^T(v - y) \end{aligned}$$

this holds for all v; therefore, $x \in \partial f^*(y)$

reverse implication $x \in \partial f^*(y) \Longrightarrow y \in \partial f(x)$ follows from $f^{**} = f$

Examples

- **Indicator function**: if $f(x) = I_C(x)$, then its conjugate function is support function of C

$$f^*(y) = I_C^*(y) = \max_{x \in C} y^T x$$

- **Norm**: if $f(x) = \|x\|$, then its conjugate is $f^*(y) = I_{\{z : \|z\|_* \leq 1\}}(y)$, where $\|\cdot\|_*$ is dual norm of $\|\cdot\|$

Why? Note that if $\|y\|_* > 1$, then there exists $\|z\| \leq 1$ with $z^T y = \|y\|_* > 1$, so

$$(tz)^T y - \|tz\| = t(z^T y - \|z\|) \to \infty, \quad \text{as } t \to \infty$$

i.e., $f^*(y) = \infty$

On the other hand, if $\|y\|_* \leq 1$, then

$$z^T y - \|z\| \leq \|z\|\|y\|_* - \|z\| \leq 0$$

and $= 0$ when $z = 0$, so $f^*(y) = 0$

- **Matrix Logarithm**: if $f(X) = -\log\det X \left(\text{dom } f = \mathbf{S}_{++}^n\right)$, then its conjugate is $f^*(Y) = -\log\det(-Y) - n$ (From B & V page 92)

4. Operations that preserve convexity

I. Nonnegative weighted sums

II. Composition with an affine mapping

III. Point-wise maximum and supremum

5. Log-concave and Log-convex functions

Definition. A function $f(x) : \mathbb{R}^n \to \mathbb{R}$, with convex domain and $f(x) > 0$ for all $x \in \text{dom} f$, is log-convex for all $x, y \in \text{dom} f$ and $0 \leq \theta \leq 1$, we have

$$f(\theta x + (1-\theta)y) \leq f(x)^{\theta} f(y)^{1-\theta}.$$

Twice differentiable condition: $f(x)\nabla^2 f(x) \leq \nabla f(x)\nabla f(x)^T$

- The CDF of a Gaussian distribution is Log-concave.
- Gamma function is Log-convex
- The Wishart density is Log-concave

6. Convex Optimization Problems

A convex optimization problem is

$$\begin{aligned} &\min_{x \in D} f(x) \\ &\text{subject to } g_i(x) \leq 0, i = 1, \ldots m \\ &Ax = b \end{aligned}$$

if criterion f is strictly convex, then the solution is unique.

The problem can be rewritten as

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

where C is feasible set.

For a convex problem and differentiable f, the first order optimal condition is

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

E.g., if we decompose $x = (x_1, x_2) \in \mathbb{R}^{n_1+n_2}$, then

$$\begin{array}{ll} \min\limits_{x_1,x_2} & f(x_1,x_2) \\ \text{s.t.} & g_1(x_1) \leq 0 \\ & g_2(x_2) \leq 0 \end{array} \iff \begin{array}{ll} \min\limits_{x_1} & \tilde{f}(x_1) \\ \text{s.t.} & g_1(x_1) \leq 0 \end{array}$$

where $\tilde{f}(x_1) = \min\{f(x_1,x_2) : g_2(x_2) \leq 0\}$. The right problem is convex if the left problem is

Partial Optimization

$g(x) = \min\limits_{y \in C} f(x,y)$ is convex in x, provided that f is convex in (x, y) and C is a convex set.

Hinge form of SVMs

Recall the SVM problem

$$\begin{array}{ll} \min\limits_{\beta,\beta_0,\xi} & \frac{1}{2}\|\beta\|_2^2 + C\sum\limits_{i=1}^n \xi_i \\ \text{subject to} & \xi_i \geq 0,\ y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i,\ i = 1,\ldots n \end{array}$$

Rewrite the constraints as $\xi_i \geq \max\{0, 1 - y_i(x_i^T\beta + \beta_0)\}$. Indeed we can argue that we have $=$ at solution

Therefore plugging in for optimal ξ gives the hinge form of SVMs:

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \left[1 - y_i(x_i^T\beta + \beta_0)\right]_+$$

where $a_+ = \max\{0, a\}$ is called the hinge function

Transformations and change of variables

If $h : \mathbb{R}^n \to \mathbb{R}$ is a monotone increasing transformation, then $\min f(x)$ is equivalent to $\min f(h(x))$.

If $\phi : \mathbb{R}^n \to \mathbb{R}^m$ is one-to-one, and its image covers feasible set C, then we can change variable in an optimization problem: $\min f(x)$ is equivalent to $\min f(\phi(y))$ for $\phi(y) \in C$.

Gradient descent for unconstraint problem

Gradient descent method: $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f\left(\boldsymbol{x}^t\right)$

1. Quadratic optimization

Consider the following problem

$$\text{minimize}_{\,x} \quad f(\boldsymbol{x}) := \frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{x}^*\right)^\top \boldsymbol{Q}\left(\boldsymbol{x} - \boldsymbol{x}^*\right),$$

for some $n \times n$ matrix $Q \succ 0$, where $\nabla f(\boldsymbol{x}) = \boldsymbol{Q}\left(\boldsymbol{x} - \boldsymbol{x}^*\right)$.

A. Constant stepwise rule

If $\eta_t \equiv \eta = \dfrac{2}{\lambda_1(\boldsymbol{Q}) + \lambda_n(\boldsymbol{Q})}$, then

$$\left\| x^t - x^* \right\|_2 \leq \left(\frac{\lambda_1(\boldsymbol{Q}) - \lambda_n(\boldsymbol{Q})}{\lambda_1(\boldsymbol{Q}) + \lambda_n(\boldsymbol{Q})}\right)^t \left\| \boldsymbol{x}^0 - \boldsymbol{x}^* \right\|_2.$$

B. Exact line search

If $\eta_t = \arg\min_{\eta \geq 0} f\left(\boldsymbol{x}^t - \eta \nabla f\left(\boldsymbol{x}^t\right)\right)$, then

$$f\left(\boldsymbol{x}^t\right) - f\left(\boldsymbol{x}^*\right) \leq \left(\frac{\lambda_1(\boldsymbol{Q}) - \lambda_n(\boldsymbol{Q})}{\lambda_1(\boldsymbol{Q}) + \lambda_n(\boldsymbol{Q})}\right)^{2t} \left(f\left(\boldsymbol{x}^0\right) - f\left(\boldsymbol{x}^*\right)\right).$$

2. Strongly convex and smooth problems

Strongly convex

I. $$f(\boldsymbol{y}) \geq \underbrace{f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x})}_{\text{first-order Taylor expansion}} + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2, \quad \forall \boldsymbol{x}, \boldsymbol{y}$$

II. For all x and y and all $0 \leq \lambda \leq 1$,

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) - \frac{\mu}{2}\lambda(1-\lambda)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2.$$

III. $$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle \geq \mu \|\boldsymbol{x}-\boldsymbol{y}\|_2^2, \quad \forall \boldsymbol{x}, \boldsymbol{y}$$

L - smooth

I. $$f(\boldsymbol{y}) \leq \underbrace{f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\top}(\boldsymbol{y}-\boldsymbol{x})}_{\text{first-order Taylor expansion}} + \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2, \quad \forall \boldsymbol{x}, \boldsymbol{y}.$$

II. For all x and y and all $0 \leq \lambda \leq 1$,

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \geq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) - \frac{L}{2}\lambda(1-\lambda)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2.$$

III. $$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle \geq \frac{1}{L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2, \quad \forall \boldsymbol{x}, \boldsymbol{y}.$$

IV. $$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x}-\boldsymbol{y}\|_2, \forall \boldsymbol{x}, \boldsymbol{y}.$$ (L - Lipschitz gradient)

Theorem 2.0 Let $f : \mathbb{R}^n \to \mathbb{R}$ be a m-strong convex and L-smooth function. Then for any $x, y \in \mathbb{R}^n$,

$$\langle \nabla f(x) - \nabla f(y), x-y\rangle \geq \frac{mL}{m+L}\|x-y\|_2^2 + \frac{1}{m+L}\|\nabla f(x) - \nabla f(y)\|_2^2.$$

Theorem 2.1 (GD for strongly convex and smooth functions)

Let f be μ-strongly convex and L-smooth. If $\eta_t \equiv \eta = \frac{2}{\mu+L}$, then

$$\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2 \leq \left(\frac{\kappa-1}{\kappa+1}\right)^t \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2,$$

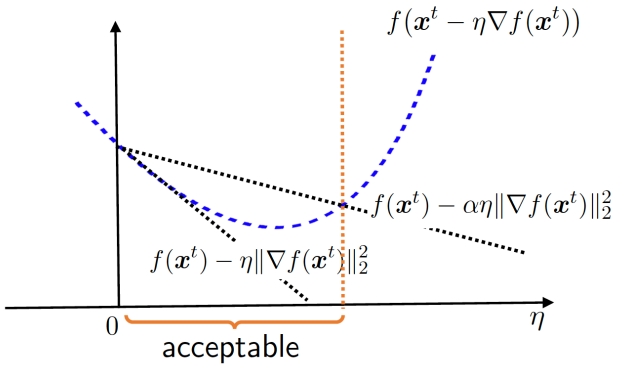
where $\kappa := L/\mu$ is condition number; $\boldsymbol{x}^*$ is minimizer

Remark.

- Dimension-free: iteration complexity is $O\left(\frac{\log \frac{1}{\varepsilon}}{\log \frac{\kappa+1}{\kappa-1}}\right)$, which is independent of problem size n if κ does not depend on n.
- Direct consequence from smoothness

$$f\left(\boldsymbol{x}^t\right)-f\left(\boldsymbol{x}^*\right) \leq \frac{L}{2}\left(\frac{\kappa-1}{\kappa+1}\right)^{2t}\left\|\boldsymbol{x}^0-\boldsymbol{x}^*\right\|_2^2.$$

Backtracking line search



Algorithm 2.2 Backtracking line search for GD

1: Initialize $\eta=1$, $0<\alpha\leq 1/2$, $0<\beta<1$
2: **while** $f(\boldsymbol{x}^t-\eta\nabla f(\boldsymbol{x}^t))>f(\boldsymbol{x}^t)-\alpha\eta\|\nabla f(\boldsymbol{x}^t)\|_2^2$ **do**
3: $\quad\eta\leftarrow\beta\eta$

Remark.

- Armijo condition: for some $0<\alpha<1$,

$$f\left(\boldsymbol{x}^t-\eta\nabla f\left(\boldsymbol{x}^t\right)\right)<f\left(\boldsymbol{x}^t\right)-\alpha\eta\left\|\nabla f\left(\boldsymbol{x}^t\right)\right\|_2^2.$$

- Ensures sufficient decrease in objective function
- Practically, backtracking line search often provides good estimate on local Lipschitz constant of gradients.

3. Other conditions

Strong convexity requirement can be relaxed

- Local strong convexity

Theorem 2.3 (GD for locally strongly convex and smooth functions)

Let f be locally μ-strongly convex and L-smooth such that

$$\mu \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq L \boldsymbol{I}, \quad \forall \boldsymbol{x} \in \mathcal{B}_0$$

where $\mathcal{B}_0 := \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \leq \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2\}$. Then Theorem 2.1 continues to hold

- Regularity condition

Another way to replace strong convexity and smoothness is regularity condition:

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq \frac{\mu}{2} \left\| \boldsymbol{x} - \boldsymbol{x}^* \right\|_2^2 + \frac{1}{2L} \|\nabla f(\boldsymbol{x})\|_2^2, \quad \forall \boldsymbol{x}.$$

Note that $\nabla f(x^*) = 0$, thus

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq \frac{\mu}{2} \left\| \boldsymbol{x} - \boldsymbol{x}^* \right\|_2^2 + \frac{1}{2L} \left\| \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}^*) \right\|_2^2, \quad \forall \boldsymbol{x}.$$

Theorem 2.4

Suppose f satisfies (2.7). If $\eta_t \equiv \eta = \frac{1}{L}$, then

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

- Polyak-Lojasiewicz condition

$$\|\nabla f(\boldsymbol{x})\|_2^2 \geq 2\mu \left(f(\boldsymbol{x}) - f(\boldsymbol{x}^*)\right), \quad \forall \boldsymbol{x}.$$

Theorem 2.5

Suppose f satisfies (2.8). If $\eta_t \equiv \eta = \frac{1}{L}$, then

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) \leq \left(1 - \frac{\mu}{L}\right)^t \left(f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)\right)$$

4. Dropping strong convexity

$f(x)$ is convex and L-smooth Key idea: majorization-minimization

Find a simple majoring function of $f(x)$ and optimize it. From smooth assumption,

$$\begin{aligned} f\left(\boldsymbol{x}^{t+1}\right) - f\left(\boldsymbol{x}^t\right) &\leq \nabla f\left(\boldsymbol{x}^t\right)^{\top}\left(\boldsymbol{x}^{t+1} - \boldsymbol{x}^t\right) + \frac{L}{2}\left\| \boldsymbol{x}^{t+1} - \boldsymbol{x}^t \right\|_2^2 \\ &= -\eta_t \left\| \nabla f\left(\boldsymbol{x}^t\right) \right\|_2^2 + \frac{\eta_t^2 L}{2}\left\| \nabla f\left(\boldsymbol{x}^t\right) \right\|_2^2. \\ &\leq -\frac{1}{2L}\left\| \nabla f\left(\boldsymbol{x}^t\right) \right\|_2^2 \end{aligned}$$

Fact 2.8

Let f be convex and L-smooth. If $\eta_t \equiv \eta = 1/L$, then

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2 \leq \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2 - \frac{1}{L^2}\|\nabla f(\boldsymbol{x}^t)\|_2^2$$

where $\boldsymbol{x}^*$ is any minimizer with optimal $f(\boldsymbol{x}^*)$

One can further show $\left\| \boldsymbol{x}^t - \boldsymbol{x}^* \right\|_2$ is strictly decreasing unless x^* is already minimizer.

Theorem 2.9 (GD for convex and smooth problems)

Let f be convex and L-smooth. If $\eta_t \equiv \eta = 1/L$, then GD obeys

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) \leq \frac{2L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}{t}$$

where $\boldsymbol{x}^*$ is any minimizer with optimal $f(\boldsymbol{x}^*)$

Generalized Steepest descent

Let p, q be complementary (dual): $1/p + 1/q = 1$

Steepest descent updates are $x^+ = x + t \cdot \Delta x$, where

$$\Delta x = \|\nabla f(x)\|_q \cdot$$
$$u = \text{argmin}_{\|v\|_p \leq 1} \nabla f(x)^T v$$

- If $p = 2$, then $\Delta x = -\nabla f(x)$, gradient descent
- If $p = 1$, then $\Delta x = -\partial f(x)/\partial x_i \cdot e_i$, where

$$\left|\frac{\partial f}{\partial x_i}(x)\right| = \max_{j=1,\ldots n} \left|\frac{\partial f}{\partial x_j}(x)\right| = \|\nabla f(x)\|_\infty.$$

Normalized steepest descent just takes $\nabla x = u$, with respect to l_1 norm: updates are

$$x_i^+ = x_i - t \cdot \text{sign}\left(\frac{\partial f}{\partial x_i}(x)\right).$$

First order method

Iterative method, updates $x^{(k)}$ in $x^{(0)} + \text{span}\left\{\nabla f\left(x^{(0)}\right), \nabla f\left(x^{(1)}\right), \ldots \nabla f\left(x^{(k-1)}\right)\right\}$.

Theorem (Nesterov): For any $k \leq (n-1)/2$ and any starting point $x^{(0)}$, there is a function f in the problem class such that any first-order method satisfies

$$f(x^{(k)}) - f^\star \geq \frac{3L\|x^{(0)} - x^\star\|_2^2}{32(k+1)^2}$$

5. 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 $g^{(k-1)} \in \partial f\left(x^{(k-1)}\right)$, any subgradients 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.,

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

Step size choices

- Fixed step size: $t_k = t$ all $k = 1,2,3,\ldots$
- Diminishing step size: $\sum_{k=1}^{\infty} t_k^2 < \infty, \quad \sum_{k=1}^{\infty} t_k = \infty$

Important difference to gradient descent: all step sizes options are pre-specified, not adaptively computed.

Convergence analysis

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

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

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

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

Theorem: For diminishing step sizes, subgradient method satisfies

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

Can prove both results form same basic inequality. Key steps:

- Using the definition of subgradient

$$\left\| x^{(k)} - x^\star \right\|_2^2 \leq \left\| x^{(k-1)} - x^\star \right\|_2^2 - 2t_k \left(f\left(x^{(k-1)}\right) - f\left(x^\star\right) \right) + t_k^2 \left\| g^{(k-1)} \right\|_2^2$$

- Iterating last inequality

$$\left\| x^{(k)} - x^\star \right\|_2^2 \leq \left\| x^{(0)} - x^\star \right\|_2^2 - 2\sum_{i=1}^{k} t_i \left(f\left(x^{(i-1)}\right) - f\left(x^\star\right) \right) + \sum_{i=1}^{k} t_i^2 \left\| g^{(i-1)} \right\|_2^2$$

- Using $\left\| x^{(k)} - x^\star \right\|_2 \geq 0$ and letting $R = \left\| x^{(0)} - x^\star \right\|_2$,

$$0 \leq R^2 - 2\sum_{i=1}^{k} t_i \left(f\left(x^{(i-1)}\right) - f\left(x^\star\right) \right) + G^2 \sum_{i=1}^{k} t_i^2$$

- Introducing $f\left(x_{\text{best}}^{(k)}\right) = \min_{i=0,\ldots k} f\left(x^{(i)}\right)$, and rearranging,

$$f\left(x_{\text{best}}^{(k)}\right) - f\left(x^\star\right) \leq \frac{R^2 + G^2 \sum_{i=1}^{k} t_i^2}{2\sum_{i=1}^{k} t_i}$$

We call this basic inequality.

After k steps with fixed step size t, basic inequality gives

$$f\left(x_{\text{best}}^{(k)}\right) - f^\star \leq \frac{R^2}{2kt} + \frac{G^2 t}{2}.$$

Proximal gradient descent

1. Decomposable functions

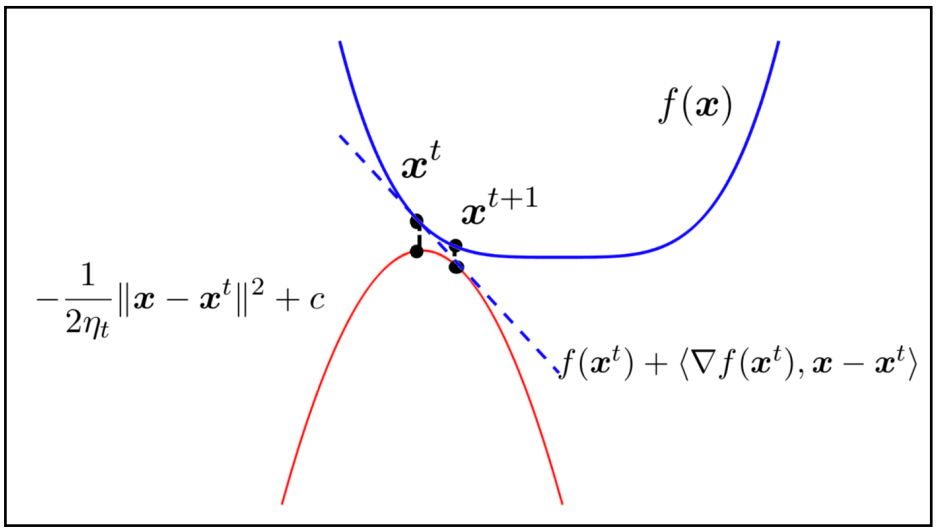
Suppose $f(x) = g(x) + h(x)$, where

- g is convex, differentiable, dom $g = \mathbb{R}^n$
- h is convex, not necessarily differentiable

If f were differentiable, minimize quadratic approximation to f around x, replace $\nabla^2 f(x)$ by $\frac{1}{t} I$,

$$x^+ = \underset{z}{\operatorname{argmin}} \underbrace{f(x) + \nabla f(x)^T (z - x) + \frac{1}{2t}\|z - x\|_2^2}_{\tilde{f}(z)},$$

which is gradient descent.



In our case, f isn't differentiable, but g is differentiable. Idea: make quadratic approximation to g, leave h alone.

$$
\begin{aligned}
x^+ &= \underset{z}{\operatorname{argmin}}\, \widetilde{g}_t(z) + h(z) \\
&= \underset{z}{\operatorname{argmin}}\, g(x) + \nabla g(x)^T(z-x) + \frac{1}{2t}\|z-x\|_2^2 + h(z) \\
&= \underset{z}{\operatorname{argmin}}\, \frac{1}{2t}\|z - (x - t\nabla g(x))\|_2^2 + h(z)
\end{aligned}
$$

stay close to gradient update for g, also makes h small.

2. *Proximal gradient descent*

Define proximal mapping:

$$
\text{prox}_t(x) = \underset{z}{\operatorname{argmin}}\, \frac{1}{2t}\|x-z\|_2^2 + h(z).
$$

Proximal gradient descent: choose initialize $x^{(0)}$, repeat

$$
x^{(k)} = \text{prox}_{t_k}\left(x^{(k-1)} - t_k \nabla g\left(x^{(k-1)}\right)\right), \quad k = 1,2,3,\ldots.
$$

To make this update step look familiar, can rewrite it as $x^{(k)} = x^{(k-1)} - t_k \cdot G_{t_k}\left(x^{(k-1)}\right)$, where G_t is the generalized gradient of f

$$
G_t(x) = \frac{x - \text{prox}_t(x - t\nabla g(x))}{t}.
$$

With criterion $f(x) = g(x) + h(x)$, we assume

- g is convex, differentiable, dom(g) = R^n, and ∇g is Lipschitz continuous with constant L > 0
- h is convex, $\text{prox}_t(x) = \text{argmin}_z \left\{\|x-z\|_2^2/(2t) + h(z)\right\}$ can be evaluated

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

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

Backtracking line search

Recall that for unconstrained case, backtracking line search is based on sufficient decrease criterion

$$f\left(\boldsymbol{x}^t - \eta \nabla f\left(\boldsymbol{x}^t\right)\right) \leq f\left(\boldsymbol{x}^t\right) - \frac{\eta}{2} \left\| \nabla f\left(\boldsymbol{x}^t\right) \right\|_2^2,$$

where we set $\alpha = 1/2$. As a result, this is equivalent to updating $\eta_t = 1/L_t$ until

$$\begin{aligned} f\left(\boldsymbol{x}^t - \eta_t \nabla f\left(\boldsymbol{x}^t\right)\right) &\leq f\left(\boldsymbol{x}^t\right) - \frac{1}{L_t}\left\langle \nabla f\left(\boldsymbol{x}^t\right), \nabla f\left(\boldsymbol{x}^t\right)\right\rangle + \frac{1}{2L_t} \left\| \nabla f\left(\boldsymbol{x}^t\right) \right\|_2^2 \\ &= f\left(\boldsymbol{x}^t\right) - \left\langle \nabla f\left(\boldsymbol{x}^t\right), \boldsymbol{x}^t - \boldsymbol{x}^{t+1}\right\rangle + \frac{L_t}{2} \left\| \boldsymbol{x}^t - \boldsymbol{x}^{t+1} \right\|_2^2 \end{aligned}.$$

Similar to gradient descent, but operates on g and not f. We fix a parameter $0 < \beta < 1$. At each iteration, start with $t = 1$, while

$$g\left(x^+\right) > g(v) + \nabla g(v)^T\left(x^+ - v\right) + \frac{1}{2t} \left\| x^+ - v \right\|_2^2,$$

where $x^+ = \text{prox}_t(v - t\nabla g(v))$. Shrink $t = \beta t$, else keep x^+.

Special cases

A. $h = 0$, gradient descent

B. $h = I_C$, projected gradient descent

C. $g = 0$, proximal minimization algorithm

3. Acceleration

Accelerated proximal gradient method: choose an initial point $x^{(0)} = x^{(-1)} \in \mathbb{R}^n$, repeat for $k = 1,2,3...$

$$
\begin{aligned}
v &= x^{(k-1)} + \frac{k-2}{k+1}\left(x^{(k-1)} - x^{(k-2)}\right) \\
x^{(k)} &= \text{prox}_{t_k}\left(v - t_k \nabla g(v)\right)
\end{aligned}
.
$$

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

$$
f(x^{(k)}) - f^\star \leq \frac{2\|x^{(0)} - x^\star\|_2^2}{t(k+1)^2}
$$

4. LASSO

Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, lasso problem can be parametrized as:

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

where $\lambda > 0$.

A. Sub-gradient method

update: $\beta^{(k)} = \beta^{(k-1)} + X^T(y - X\beta^{(k-1)}) - \partial\|\beta^{(k-1)}\|_1$

B. Proximal gradient descent (ISTA)

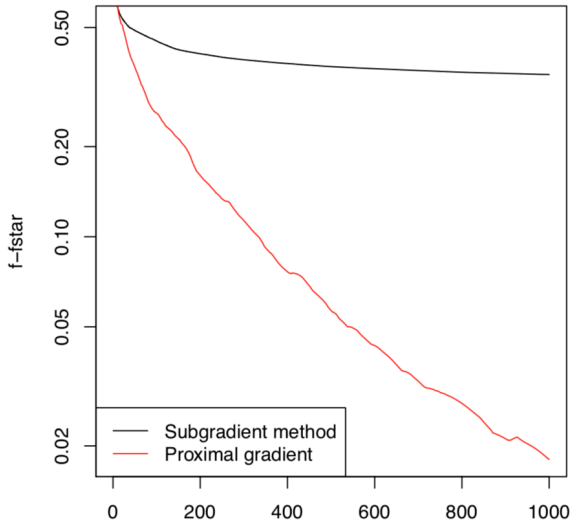
Proximal mapping is now

$$
\begin{aligned}
\text{prox}_t(\beta) &= \underset{z \in \mathbb{R}^n}{\text{argmin}} \frac{1}{2t}\|\beta - z\|_2^2 + \lambda\|z\|_1 \\
&= S_{\lambda t}(\beta)
\end{aligned}
,
$$

Where $S_\lambda(\beta)$ is the soft-thresholding operator,

$$
[S_\lambda(\beta)]_i = \begin{cases} \beta_i - \lambda & \text{if } \beta_i > \lambda \\ 0 & \text{if } -\lambda \leq \beta_i \leq \lambda \\ \beta_i + \lambda & \text{if } \beta_i < -\lambda \end{cases}, \quad i = 1, \ldots n
$$

Proximal gradient update is: $\beta^+ = S_{\lambda t}\left(\beta + tX^T(y - X\beta)\right)$, which is often called iterative soft-thresholding algorithm.


f−fstar
0.50
0.20
0.10
0.05
0.02
0
200
400
600
800
1000
Subgradient method
Proximal gradient

Duality in General Programs

1. Lagrangian

Consider general minimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{subject to} & h_i(x) \leq 0,\ i = 1,2,...,m \\ & \ell_j(x) = 0,\ j = 1,2,...,m \end{array}$$

We define the Lagrangian as

$$L(x,u,v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x),$$

new variables $u \in \mathbb{R}^m, v \in \mathbb{R}^r$, with $u \geq 0$.

Important property: for any $u \geq 0$ and v,

$$f(x) \geq L(x,u,v) \text{ at each feasible } x.$$

2. Lagrange dual problem

Let C denote the feasible set, $f^\star$ denote primal optimal value, then

$$f^\star \geq \min_{x \in C} L(x,u,v) \geq \min_{x} L(x,u,v) := g(u,v).$$

We call $g(u,v)$ the Lagrange dual function, and it gives a lower bound on $f^\star$ for any $u \geq 0$ and v, called feasible dual u, v.

Best lower bound is given by maximizing $g(u,v)$ over all feasible dual u, v, yielding Lagrange dual problem:

$$\begin{array}{ll} \max\limits_{u,v} & g(u,v) \\ \text{subject to} & u \geq 0 \end{array}.$$

Important properties:

- Dual problem is always convex
- The primal optimal value $f^\star$ and dual optimal $g^\star$ satisfies weak duality: $f^\star \geq g^\star$

- Slater's condition: for convex primal, if there is an x such that

$$h_1(x) < 0, \ldots h_m(x) < 0 \quad \text{and} \quad \ell_1(x) = 0, \ldots \ell_r(x) = 0$$

Then strong duality holds: $f^\star = g^\star$.

3. KKT conditions

For a problem with strong duality, $x^\star$ and $u^\star$, $v^\star$ are primal and dual solutions $\iff$ $x^\star$ and $u^\star$, $v^\star$ satisfy the KKT conditions. The Karush-Kuhn-Tucker conditions are:

- Stationarity $0 \in \partial f(x) + \sum_{i=1}^{m} u_i \partial h_i(x) + \sum_{j=1}^{r} v_j \partial \ell_j(x)$
- Complementary slackness $u_i \cdot h_i(x) = 0$ for all i
- Primal feasibility $h_i(x) \leq 0$, $l_i(x) = 0$ for all i, j
- Dual feasibility $u_i \geq 0$ for all i

Theorem: Let f be differentiable and strictly convex, let $X \in \mathbb{R}^{n \times p}$, $\lambda > 0$. Consider

$$\min_{\beta \in \mathbb{R}^p} \; f(X\beta) + \lambda \|\beta\|_1$$

If the entries of X are drawn from a continuous probability distribution (on $\mathbb{R}^{np}$), then w.p. 1 there is a unique solution and it has at most $\min\{n, p\}$ nonzero components

4. Conjugates and dual problem

Consider

$$\min_x f(x) + g(x)$$
$$\iff \min_{x,z} f(x) + g(z) \text{ subject to } x = z$$

Lagrange dual function

$$g(u) = \min_x f(x) + g(z) + u^T(z - x) = -f^*(u) - g^*(-u),$$

Hence dual problem is $\max_u -f^*(u) - g^*(-u)$

Example: Lasso dual

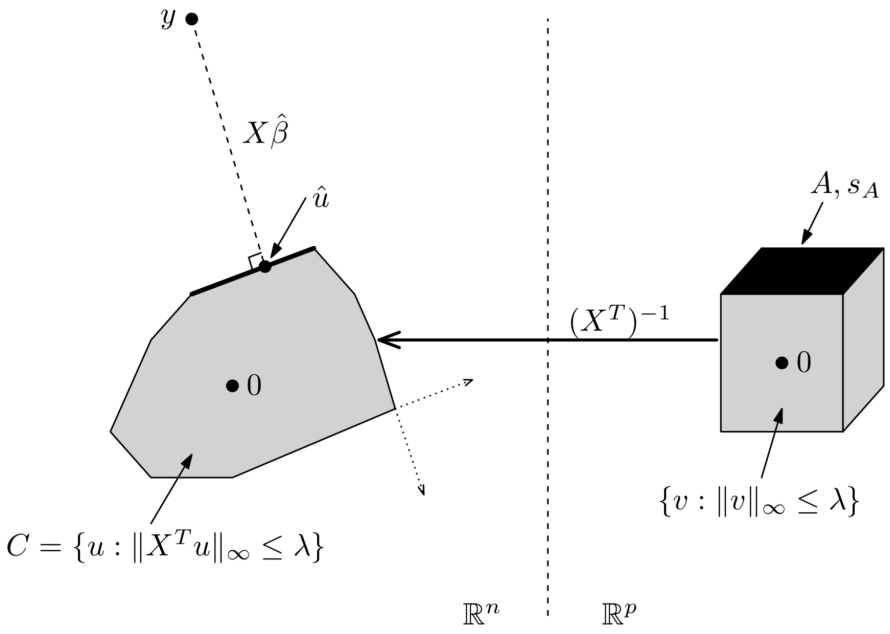
Primal problem is

$$\min_{\beta\in\mathbb{R}^p, z\in\mathbb{R}^n} \frac{1}{2}\|y-z\|_2^2 + \lambda\|\beta\|_1 \text{ subject to } z = X\beta,$$

So dual function is now

$$\begin{aligned} g(u) &= \min_{\beta\in\mathbb{R}^p, z\in\mathbb{R}^n} \frac{1}{2}\|y-z\|_2^2 + \lambda\|\beta\|_1 + u^T(z - X\beta) \\ &= \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y-u\|_2^2 - I_{\{v:\|v\|_\infty\le 1\}}\left(X^T u/\lambda\right) \end{aligned}.$$

Therefore Lasso dual problem is $\min_{u\in\mathbb{R}^n} \|y-u\|_2^2$ subject to $\left\| X^T u \right\|_\infty \le \lambda$. Further, note that given the dual solution u, any Lasso solution satisfies $X\beta = y - u$. This is from the KKT stationary condition for z (i.e. $z - y + u = 0$).



5. *Dual cones and dual problems*

Dual cones: for a cone $K \subseteq \mathbb{R}^n$ (recall this means $x \in K, t \ge 0 \Longrightarrow tx \in K$) and the dual cone is

$$K^* = \left\{y \in \mathbb{R}^n : y^T x \ge 0 \text{ for all } x \in K\right\}.$$

Examples

- **Linear subspace**: the dual cone of linear subspace K is $K^{\perp}$
- **Norm cone**: the dual cone of the norm cone $K = \left\{(x,t) \in \mathbb{R}^{n+1} : \|x\| \leq t\right\}$ is the norm cone its dual norm $K^* = \left\{(y,s) \in \mathbb{R}^{n+1} : \|y\|_* \leq s\right\}$

To prove the result we have to show

$$x^T y + ts \geq 0 \text{ whenever } \|x\| \leq t \iff \|y\|_* \leq s.$$

($\Longleftarrow$) Suppose $\|y\|_* \leq s$ and $\|x\| \leq t$ for some $t > 0$ ($t = 0$ obviously hold). Applying the definition of dual norm and the fact that $\| - x/t\| \leq 1$, we have

$$y^T(-x/t) \leq \|y\|_* \leq s,$$

And therefore $x^T y + ts \geq 0$.

($\Longrightarrow$) Suppose $\|y\|_* > s$, then there exists x with $\|x\| \leq 1$ and $x^T y \geq s$. Taking $t = 1$, we have

$$-x^T y + s \leq 0.$$

- **Positive semidefinite cone**: the convex cone $\mathbb{S}^n_+$ is self-dual, meaning $(\mathbb{S}^n_+)^* = \mathbb{S}^n_+$.

Next we will show $Y \succeq 0 \iff \text{tr}(YX) \geq 0$ for all $X \succeq 0$.

($\Longleftarrow$) Suppose $Y \not\succeq 0$, then there exists an $u \in \mathbb{R}^n$ we have $u^T Y u = tr(uu^T Y) < 0$, and let $X = uu^T$.

($\Longrightarrow$) The eigenvalue decomposition of X is $X = Q\Lambda Q^T = \sum_{i=1}^{n} \lambda_i q_i q_i^T$, then we have

$$tr(YX) = tr(\sum_{i=1}^{n} \lambda_i q_i^T Y q_i) \geq 0.$$

Consider the cone constrained problem $\min_{x \in K} f(x)$. Recall that its dual problem is $\max_{u \in \mathbb{R}^n} -f^*(u) - I_K^*(-u)$, where $I_K^*(-u) = I_{K^*}(u)$, and this is simply $\max_{x \in K^*} -f^*(x)$.

6. SVM

Functional and geometric margins

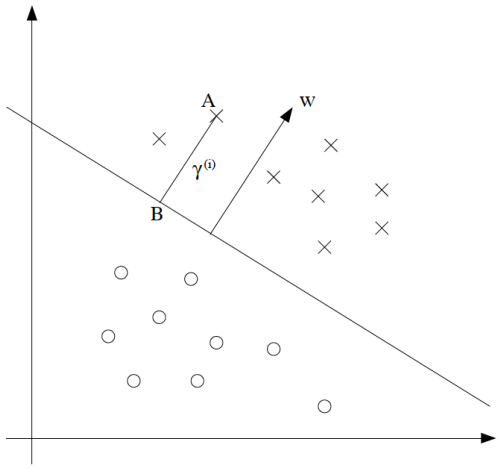
Given a training example $(x^{(i)}, y^{(i)})$, we define the function margin of (ω, b) with respect to the training example

$$\hat{\gamma}^{(i)} = y^{(i)}\left(\omega^T x + b\right).$$

Given a training set $S = \{(x^{(i)}, y^{(i)}) : i = 1,2,...,m\}$, we also define the function margin of (ω, b) with respect to the training set

$$\hat{\gamma} = \min_i \hat{\gamma}^{(i)}.$$

Now let's talk about geometric margins, $\gamma^{(i)}$ is length of line segment AB, i.e., the distance from a training sample to the line $\omega^T x + b = 0$.



Since A represents $x^{(i)}$, therefore B is given by $x^{(i)} - \gamma^{(i)}\omega/\|\omega\|$. And point B lies on the decision boundary. Hence,

$$w^T\left(x^{(i)} - \gamma^{(i)}\frac{w}{\|w\|}\right) + b = 0,$$

Solving for $\gamma^{(i)}$ yields

$$\gamma^{(i)} = \frac{w^T x^{(i)} + b}{\|w\|} = \left(\frac{w}{\|w\|}\right)^T x^{(i)} + \frac{b}{\|w\|}.$$

Note that if $\|\omega\| = 1$, the geometric margin equals to functional marginal.

The optimal margin classifier

We can pose the following optimization problem:

$$\begin{aligned} &\max_{\hat{\gamma},\omega,b} \hat{\gamma} \\ &s.t.\, y^{(i)}\left(\omega^T x^{(i)} + b\right) \geq \hat{\gamma} \\ &\quad \|\omega\| = 1 \end{aligned}.$$

Let's try transforming the problem into a nicer one, consider

$$\begin{aligned} &\max_{\hat{\gamma},\omega,b} \frac{\hat{\gamma}}{\|\omega\|} \\ &s.t.\, y^{(i)}\left(\omega^T x^{(i)} + b\right) \geq \hat{\gamma} \end{aligned}.$$

We will introduce the scaling constraint that the functional margin of w, b with respect to the training set must be 1: $\hat{\gamma} = 1$. Then we can transform the optimization problem to

$$\begin{aligned} &\min_{\omega,b} \frac{1}{2}\|\omega\|^2 \\ &s.t.\, y^{(i)}\left(\omega^T x^{(i)} + b\right) \geq 1 \end{aligned}.$$

Lagrange duality

The Lagrange dual function of optimization problem is

$$\mathscr{L}(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}\left(w^T x^{(i)} + b\right) - 1\right].$$

Then we obtain the following dual problem

$$\begin{aligned} \max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \left\langle x^{(i)}, x^{(j)} \right\rangle \\ & \alpha_i \geq 0, i = 1,2,...,m \\ & \sum_{i=1}^{n} \alpha_i y^{(i)} = 0 \end{aligned}.$$

Newton's method

Given unconstrained, smooth and convex optimization $\min_x f(x)$, where f is convex, twice differentiable, and $\text{dom}\, f = \mathbb{R}^n$. Newton's method repeats

$$x^{(k)} = x^{(k-1)} - \left(\nabla^2 f\left(x^{(k-1)}\right)\right)^{-1} \nabla f\left(x^{(k-1)}\right), \quad k = 1,2,3,\ldots.$$

Newton's method interpretation

- Better quadratic approximation

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

And minimize over y to yield $x^+ = x - \left(\nabla^2 f(x)\right)^{-1} \nabla f(x)$.

- Linearized optimality condition $0 = \nabla f(x+v) \approx \nabla f(x) + \nabla^2 f(x)v$ and solving for v, which again yields $v = -\left(\nabla^2 f(x)\right)^{-1} \nabla f(x)$.

Affine invariance of Newton's method

Given f, nonsingular $A \in \mathbb{R}^{m\times n}$. Let $x = Ay$, and $g(y) = f(Ay)$. Newton steps on g are

$$x^+ = x - \left(\nabla^2 f(x)\right)^{-1} f(x).$$

The Newton decrement

At a point x, we define the Newton decrement $\lambda(x) = \left(\nabla f(x)^T \left(\nabla^2 f(x)\right)^{-1} \nabla f(x)\right)^{1/2}$, which is the difference between $f(x)$ and minimum of its quadratic approximation.

$$f(x) - \min_y f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x)(y-x) = \frac{1}{2}\lambda^2(x).$$

Backtracking line search

In practice, we use damped Newton's method, which repeats $x^+ = x - t\left(\nabla^2 f(x)\right)^{-1} f(x)$. Steps size here typically are chosen by backtracking line search, with parameters $0 < \alpha < 1/2$, $0 < \beta < 1$. At each iteration, we start with $t = 1$ and while

$$f(x + tv) > f(x) + \alpha t \nabla f(x)^T v,$$

We shrink $t = \beta t$, else we perform Newton update. Note that here $v = -\left(\nabla^2 f(x)\right)^{-1} \nabla f(x)$, thus $\nabla f(x)^T v = -\lambda^2(x)$ is a constant.

Convergence analysis

Assume that f convex, twice differentiable, having $\text{dom}\, f = \mathbb{R}^n$ and additionally,

- $\nabla f(x)$ is Lipschitz with parameter L
- f is strongly convex with parameter m
- $\nabla^2 f(x)$ is Lipschitz with parameter M

Theorem: Newton's method with backtracking line search satisfies the following two-stage convergence bounds

$$f(x^{(k)}) - f^\star \leq \begin{cases} (f(x^{(0)}) - f^\star) - \gamma k & \text{if } k \leq k_0 \\ \dfrac{2m^3}{M^2}\left(\dfrac{1}{2}\right)^{2^{k-k_0+1}} & \text{if } k > k_0 \end{cases}$$

Here $\gamma = \alpha\beta^2\eta^2 m/L^2$, $\eta = \min\{1, 3(1-2\alpha)\}m^2/M$, and k_0 is the number of steps until $\|\nabla f(x^{(k_0+1)})\|_2 < \eta$

Convergence follows two stages:

- Damped phase: $\|\nabla f(x^{(k)})\|_2 \geq \eta$, and $f(x^{(k)}) - f(x^{(k+1)}) \geq \gamma$.
- Pure phase: $\|\nabla f(x^{(k)})\|_2 < \eta$, backtracking selects $t = 1$, and

$$\frac{M}{2m^2} \left\| \nabla f\left(x^{(k+1)}\right) \right\|_2 \leq \left(\frac{M}{2m^2} \left\| \nabla f\left(x^{(k)}\right) \right\|_2 \right)^2.$$

To reach $f(x^{(k)}) - f^* \leq \epsilon$, we need at most $\dfrac{f\left(x^{(0)}\right) - f^\star}{\gamma} + \log\log\left(\epsilon_0/\epsilon\right)$ iterations, where $\epsilon_0 = 2m^3/M^2$.

Strong advantages

- Convergence of Newton's method is rapid in general, and quadratic near x^*.
- Newton's method is affine invariant.

- Newton's method scales well with problem size. Its performance on problem $\mathbb{R}^{10000}$ is similar to which on problem $\mathbb{R}^{10}$.
- The good performance of Newton's method is not depend on the choice of algorithm parameters.

The main disadvantage of Newton's method is cost of **forming and storing the Hessian**, and the cost of **computing the Newton step**, which requires solving a set of linear equations.

Self-Concordance

Two major shortcoming of classical convergence analysis of Newton's method:

- The unknown parameters L, m, M in practice
- Dependent on coordinate system used

A scale-free analysis is possible for self-concordant functions: on $\mathbb{R}$, a convex function f is called self-concordant if

$$|f'''(x)| \leq 2f''(x)^{3/2}.$$

And on $\mathbb{R}^n$ is called self-concordant if its projection on every line segment is so.

Theorem (Nesterov and Nemirovskii): Newton's method with backtracking line search requires at most

$$C(\alpha, \beta)\big(f(x^{(0)}) - f^\star\big) + \log\log(1/\epsilon)$$

iterations to reach $f(x^{(k)}) - f^\star \leq \epsilon$, where $C(\alpha, \beta)$ is a constant that only depends on α, β

Comparison to first-order methods

- Memory: each iteration of Newton's method requires $O(n^2)$ storage ($n \times n$ Hessian); Each gradient iteration requires $O(n)$ storage (n-dimensional gradient).
- Computation: each Newton iteration requires $O(n^3)$ flops (solving a dense $n \times n$ linear system); Each gradient iteration requires $O(n)$ flops (scaling/adding n-dimensional gradient)

- Backtracking: both use $O(n)$ flops per inner backtracking step.

Equality-constrained Newton's method

Consider now a problem with equality constraints, as in

$$\min_x f(x) \text{ subject to } Ax = b.$$

Several options:

- Eliminating equality constraints: write $x = Fy + b$, where F spans null space of A, and $Ax_0 = b$. Solve in terms of y.
- Deriving the dual: can check the Lagrange dual function $-f^*(-A^T v) - b^T v$, and strong duality holds. With luck, we can express x^* in terms of v^*.
- Equality-constrained Newton:

In equality-constrained Newton's method we start with $x^{(0)}$ such that $Ax^{(0)} = b$. Then we repeat the updates $x^+ = x + tv$, where

$$v = \underset{Az=0}{\operatorname{argmin}} \nabla f(x)^T (z - x) + \frac{1}{2}(z - x)^T \nabla^2 f(x)(z - x).$$

Furthermore, v is the solution to minimizing a quadratic subject to equality constraints. From the KKT conditions that v satisfies

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix},$$

For some w. Hence Newton direction is given by solving linear system in Hessian.

Barrier Method

1. *Hierarchy of second-order method*

Assuming all problems are convex

- Quadratic problems are the easiest: closed form solution
- Equality-constrained quadratic problems are still easy: we use KKT conditions to derive closed form solution
- Equality-constrained smooth problems are next: use Newton's method to reduce this to a sequence of equality constrained quadratic problems
- Inequality- and equality-constrained smooth problems: use interior methods to reduce this a sequence of equality constrained quadratic problems

2. *Log barrier function*

Consider the convex optimization problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0,\ i = 1,2,...,m \cdot \\ & Ax = b \end{aligned}$$

We will assume that f, $h_1(x), \ldots, h_m(x)$ are convex and twice differentiable, each with the domain $\mathbb{R}^n$. The function

$$\phi(x) = \sum_{i=1}^{m} -\log(-h_i(x))$$

Is called the log barrier for the above problem. Its domain is the strictly feasible points, $\{x : h_i(x) < 0,\ i = 1,2,...,m\}$.

Ignoring the equality constraints for now, the problem can be written as

$$\min_x f(x) + \sum_{i=1}^{m} I(h_i(x) \leq 0).$$

We approximate this representation by adding log barrier function

$$\min_x f(x) - (1/t)\sum_{i=1}^{m} \log(-h_i(x)),$$

Where $t > 0$ is a large number.

3. *Central path*

Consider minimize our problem

$$\begin{array}{ll} \min\limits_x & tf(x) + \phi(x) \\ \text{subject to} & Ax = b \end{array}$$

The **central path** is defined as the solution $x^*(t)$ as a function of $t > 0$. These solutions are characterized by KKT conditions:

$$Ax^\star(t) = b, h_i\left(x^\star(t)\right) < 0, \quad i = 1,\ldots m,$$

$$t\nabla f\left(x^\star(t)\right) - \sum_{i=1}^{m} \frac{1}{h_i\left(x^\star(t)\right)} \nabla h_i\left(x^\star(t)\right) + A^T w = 0$$

For some $w \in \mathbb{R}^m$. As $t \to \infty$, we hope that $x^\star(t) \to x^\star$.

4. *Dual points from central path*

Given $x^\star(t)$ and corresponding w, we fine

$$u_i^\star(t) = -\frac{1}{th_i(x^\star)},\ i = 1,2,\ldots,m,\ v^\star(t) = w/t.$$

Note that $u_i^\star(t) > 0$ for all i, furthermore, the point $(u^\star(t),\ v^\star(t))$ lies in domain of Lagrange dual function $g(u, v)$, since by definition

$$\nabla f\left(x^\star(t)\right) + \sum_{i=1}^{m} u_i\left(x^\star(t)\right) \nabla h_i\left(x^\star(t)\right) + A^T v^\star(t) = 0.$$

I.e., $x^\star(t)$ minimize the Lagrangian $L(x, u^\star(t), v^\star(t))$ over x, so $g(u^\star(t), v^\star(t)) > -\infty$.

We compute

$$\begin{aligned} g\left(u^\star(t), v^\star(t)\right) &= f\left(x^\star(t)\right) + \sum_{i=1}^{m} u_i^\star(t) h_i\left(x^\star(t)\right) + v^\star(t)^T(Ax^\star(t) - b) \\ &= f\left(x^\star(t)\right) - m/t \end{aligned}$$

That is $f(x^\star(t)) - f^\star \leq m/t$, this will be very useful as a **stopping criterion**; it also confirms the fact that as $t \to \infty, x^\star(t) \to x^\star$.

5. *Barrier method*

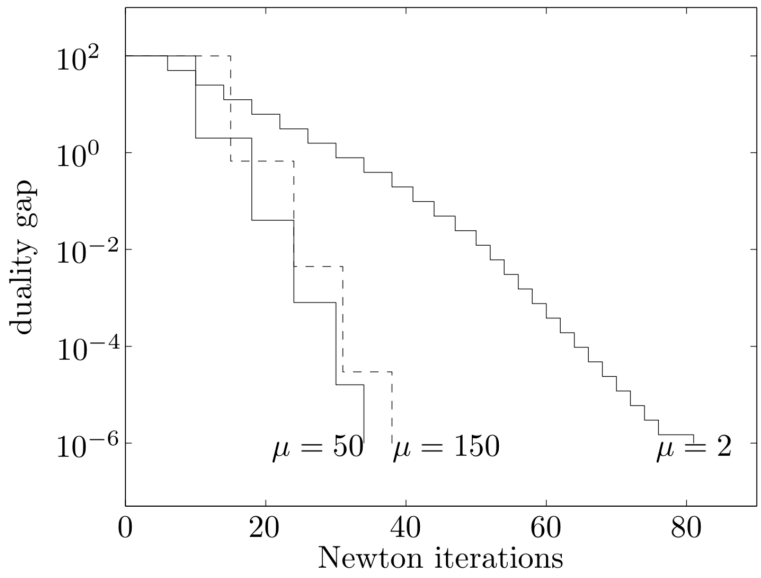
The Barrier method solves a sequence of problems

$$\begin{array}{ll} \min\limits_{x} & tf(x) + \phi(x) \\ \text{subject to} & Ax = b \end{array},$$

For increasing $t > 0$, until $m/t \leq \epsilon$. We start with $t = t^{(0)} > 0$, and solve the above problem using Newton's method to produce $x^{(0)} = x^\star(t)$. Then for barrier parameter $\mu > 1$, we repeat, for $k = 1,2,...$

- Solve the barrier problem at $t = t^{(k)}$, using Newton's method initialize at $x^{(k-1)}$, to produce $x^{(k)} = x^\star(t)$
- Stop if $m/t \leq \epsilon$
- Else update $t^{(k+1)} = \mu t$

The first step is called centering step since it brings $x^{(k)}$ onto the central path.



6. *Convergence analysis*

Theorem. *The barrier method after k centering steps satisfies* $f\left(x^{(k)}\right) - f^\star \leq \dfrac{m}{\mu^k t^{(0)}}$.

In other words, to reach a desired accuracy level of ϵ, we quire $\dfrac{\log\left(m/\left(t^{(0)}\epsilon\right)\right)}{\log\mu}+1$ centering steps with the barrier method (plus the initial centering step).

7. Feasibility method

How to find such a feasible x? By solving

$$\begin{array}{ll}\min\limits_{x,s} & s \\ \text{subject to} & h_i(x)\le s, \quad i=1,\dots m \\ & Ax=b\end{array}$$

The goal is for s to be negative at the solution. Once we find a feasible (x,s) with $s<0$, we can terminate early.

An alternative is to solve the problem

$$\begin{array}{ll}\min\limits_{x,s} & 1^T s \\ \text{subject to} & h_i(x)\le s_i, \quad i=1,\dots m \\ & Ax=b, \quad s\ge 0\end{array}$$

The nonzero entries of s will tell us which of the constraints cannot be satisfied.

Gradient descent for constraint problem

$$\begin{aligned} &\text{minimize}_x f(x) \\ &\text{subject to } x \in \mathscr{C} \end{aligned}$$

Where $f(x)$ is a convex function and $\mathscr{C}$ is a convex set.

1. Frank-Wolfe Algorithm

Algorithm 3.1 Frank-wolfe (a.k.a. conditional gradient) algorithm

1: **for** $t = 0, 1, \cdots$ **do**
2: $\quad \boldsymbol{y}^t := \arg\min_{\boldsymbol{x} \in \mathcal{C}} \langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x} \rangle$ (direction finding)
3: $\quad \boldsymbol{x}^{t+1} = (1 - \eta_t)\boldsymbol{x}^t + \eta_t \boldsymbol{y}^t$ (line search and update)

Direction finding is equivalent to

$$f\left(\boldsymbol{x}^t\right) + \left\langle \nabla f\left(\boldsymbol{x}^t\right), \boldsymbol{x} - \boldsymbol{x}^t \right\rangle,$$

which is a linear optimization over convex set. And step size is determined by line search

$$\boldsymbol{\eta}^t = \arg\min_{\boldsymbol{\eta}} f\left((1-\eta)x^t + \eta y^t\right).$$

Convergence analysis

Theorem 3.1 (Frank-Wolfe for convex and smooth problems, Jaggi '13)

Let f be convex and L-smooth. With $\eta_t = \frac{2}{t+2}$, one has

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) \leq \frac{2Ld_{\mathcal{C}}^2}{t+2}$$

where $d_{\mathcal{C}} = \sup_{\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}} \|\boldsymbol{x} - \boldsymbol{y}\|_2$

2. *Projected gradient methods*

Projected Theorem. Let $\mathscr{C}$ be convex set. Then $x_{\mathscr{C}}$ is projection of x onto $\mathscr{C}$ iff

$$\left(\boldsymbol{x}-\boldsymbol{x}_{\mathscr{C}}\right)^{\top}\left(\boldsymbol{z}-\boldsymbol{x}_{\mathscr{C}}\right) \leq 0, \quad \forall \boldsymbol{z} \in \mathscr{C}.$$

Nonexpansivness of projection. For any x and z, one has

$$\left\| \mathscr{P}_{\mathscr{C}}(\boldsymbol{x})-\mathscr{P}_{\mathscr{C}}(\boldsymbol{z}) \right\|_2 \leq \|\boldsymbol{x}-\boldsymbol{z}\|_2$$

3. *Strongly convex and smooth problem*

- x^* lies in interior of $\mathscr{C}$

Then x^* is a loco minimizer because $x^* \in \mathscr{C}$, which implies that $\nabla f(x^*) = 0$.

Theorem 3.5

Suppose $\boldsymbol{x}^* \in \mathrm{int}(\mathcal{C})$, and let f be μ-strongly convex and L-smooth. If $\eta_t = \frac{2}{\mu+L}$, then

$$\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2 \leq \left(\frac{\kappa-1}{\kappa+1}\right)^t \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2$$

where $\kappa = L/\mu$ is condition number

- Don't know whether x ∈ int(C)

  Main issue: $\nabla f(x^*)$ may not equal to 0 (x^* may not be loco minimizer).

Let $\boldsymbol{x}^+ := \mathscr{P}_{\mathscr{C}}\left(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})\right)$ and $\boldsymbol{g}_{\mathscr{C}}(\boldsymbol{x}) := L\left(\boldsymbol{x}-\boldsymbol{x}^+\right)$ and $\boldsymbol{g}_{\mathscr{C}}(x)$ generalizes $\nabla f(x)$ and obeys $\boldsymbol{g}_{\mathscr{C}}(x^*) = 0$.

When $x^* \in \mathscr{C}$, obviously $\boldsymbol{g}_{\mathscr{C}}(x^*) = 0$. And when $x^* \in \partial\mathscr{C}$, if $\boldsymbol{g}_{\mathscr{C}}(x^*) \neq 0$ then we can find x^{*+} such that $f(x^{*+}) \leq f(x^*)$, which is contradictory.

Theorem 3.7 (projected GD for strongly convex and smooth problems)

Let f be μ-strongly convex and L-smooth. If $\eta_t \equiv \eta = \frac{1}{L}$, then

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

- Slightly weaker convergence rate compared with Theorem 3.5.

4. *Convex and smooth problem*

Theorem 3.8 (projected GD for convex and smooth problems)

Let f be convex and L-smooth. If $\eta_t \equiv \eta = \frac{1}{L}$, then

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*) \leq \frac{3L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 + f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{t+1}$$

Proof.

Step 1: show cost improvement

$$f\left(\boldsymbol{x}^{t+1}\right) \leq f\left(\boldsymbol{x}^t\right) - \frac{1}{2L} \left\| \boldsymbol{g}_{\mathscr{C}}\left(\boldsymbol{x}^t\right) \right\|_2^2.$$

Step 2: connect $\left\| g_{\mathscr{C}}\left(\boldsymbol{x}^t\right) \right\|_2$ with $f(\boldsymbol{x}_t)$

$$\left\| \boldsymbol{g}_{\mathscr{C}}\left(\boldsymbol{x}^t\right) \right\|_2 \geq \frac{f\left(\boldsymbol{x}^{t+1}\right) - f\left(\boldsymbol{x}^*\right)}{\left\| \boldsymbol{x}^t - \boldsymbol{x}^* \right\|_2} \geq \frac{f\left(\boldsymbol{x}^{t+1}\right) - f\left(\boldsymbol{x}^*\right)}{\left\| \boldsymbol{x}^0 - \boldsymbol{x}^* \right\|_2}.$$

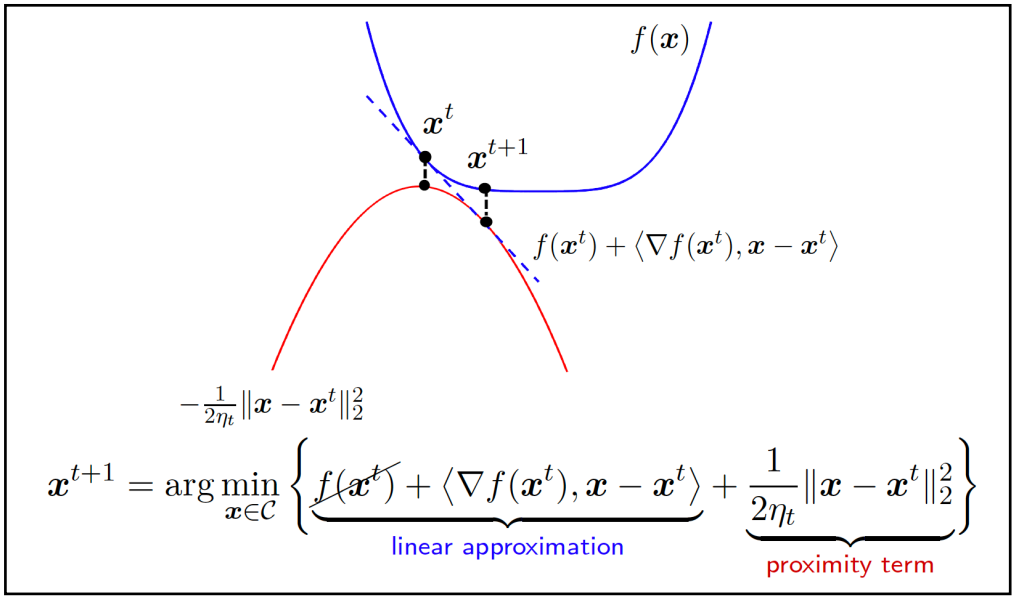
Step 3: Let $\Delta_t := f\left(\boldsymbol{x}^t\right) - f\left(\boldsymbol{x}^*\right)$ to get

$$\Delta_{t+1} - \Delta_t \leq -\frac{\Delta_{t+1}^2}{2L \left\| x^0 - x^* \right\|_2^2}.$$

And complete proof by induction.

Mirror Descent

A proximal viewpoint of projected GD



Quadratic proximal term is used to monitor discrepancy between $f(x)$ and first order approximation. (homogeneous penalty: squared Euclidean penalty)

Issues: local geometry might sometimes be highly inhomogeneous, or even non-Euclidean

Mirror descent

***Mirror descent**: adjust gradient updates to fit problem geometry*

— Nemirovski & Yudin, '1983

Replace quadratic proximity $\left\| \boldsymbol{x} - \boldsymbol{x}^t \right\|_2^2$ with distance-like metric D_φ

$$\boldsymbol{x}^{t+1} = \arg\min_{\boldsymbol{x}\in\mathscr{C}} \left\{ f\left(\boldsymbol{x}^t\right) + \left\langle \nabla f\left(\boldsymbol{x}^t\right), \boldsymbol{x} - \boldsymbol{x}^t \right\rangle + \frac{1}{\eta_t} \underbrace{D_\varphi\left(\boldsymbol{x}, \boldsymbol{x}^t\right)}_{\text{Bregman divergence}} \right\},$$

where $D_\varphi(\boldsymbol{x}, \boldsymbol{z}) := \varphi(\boldsymbol{x}) - \varphi(\boldsymbol{z}) - \langle \nabla\varphi(\boldsymbol{z}), \boldsymbol{x} - \boldsymbol{z} \rangle$ for convex and differentiable $\varphi(x)$.

Principles in choosing Bregman divergence:

- Fits local curvature of $f(x)$
- Fits geometry of constraint set $\mathscr{C}$
- Make sure Bregman projection is inexpensive

Bregman Divergence

Let $\varphi : \mathscr{C} \mapsto \mathbb{R}$ be strictly convex and differentiable on $\mathscr{C}$, then

$$D_\varphi(\boldsymbol{x}, z) := \varphi(\boldsymbol{x}) - \varphi(z) - \langle \nabla\varphi(z), \boldsymbol{x} - z\rangle.$$

- Squared Mahalanobis distance

Let $D_\varphi(\boldsymbol{x}, z) = \frac{1}{2}(\boldsymbol{x} - z)^\top \boldsymbol{Q}(\boldsymbol{x} - z)$ for $Q > 0$, which is generated by $\varphi(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}$.

- KL divergence

Let $D_\varphi(\boldsymbol{x}, z) = \mathrm{KL}(\boldsymbol{x}\|z) := \sum_i x_i \log \frac{x_i}{z_i}$, which is generated by $\varphi(\boldsymbol{x}) = \sum_i x_i \log x_i$ (negative entropy) if $\mathscr{C} = \Delta := \left\{ \boldsymbol{x} \in \mathbb{R}^n_+ \,|\, \sum_i x_i = 1 \right\}$ is a probability simplex.

Bregman Projection

Given a point $\boldsymbol{x}$, define

$$\mathscr{P}_{\mathscr{C},\varphi}(\boldsymbol{x}) := \arg\min_{z\in\mathscr{C}} D_\varphi(z, \boldsymbol{x}).$$

As Bergman projection of $\boldsymbol{x}$ onto $\mathscr{C}$.

Generalized Pythagorean Theorem. If $\boldsymbol{x}_{\mathscr{C},\varphi} = \mathscr{P}_{\mathscr{C},\varphi}(\boldsymbol{x})$, then

$$D_\varphi(z, \boldsymbol{x}) \geq D_\varphi\left(z, \boldsymbol{x}_{\mathscr{C},\varphi}\right) + D_\varphi\left(\boldsymbol{x}_{\mathscr{C},\varphi}, \boldsymbol{x}\right) \quad \forall z \in \mathscr{C}.$$

An alternative form of MD

Using Bergman divergence, one can also describe MD as

$$\nabla\varphi\left(\boldsymbol{y}^{t+1}\right) = \nabla\varphi\left(\boldsymbol{x}^t\right) - \eta_t \boldsymbol{g}^t \quad \text{with } \boldsymbol{g}^t \in \partial f\left(\boldsymbol{x}^t\right),$$

$$\boldsymbol{x}^{t+1} \in \mathscr{P}_{\mathscr{C},\varphi}\left(\boldsymbol{y}^{t+1}\right) = \arg\min_{z\in\mathscr{C}} D_{\varphi}\left(z, \boldsymbol{y}^{t+1}\right).$$

$$\begin{aligned} x_{t+1} &= \operatorname*{argmin}_{x\in\mathscr{X}\cap\mathscr{D}} D_{\Phi}\left(x, y_{t+1}\right) \\ &= \operatorname*{argmin}_{x\in\mathscr{X}\cap\mathscr{D}} \Phi(x) - \nabla\Phi\left(y_{t+1}\right)^{\top} x \\ &= \operatorname*{argmin}_{x\in\mathscr{X}\cap\mathscr{D}} \Phi(x) - \left(\nabla\Phi\left(x_t\right) - \eta g_t\right)^{\top} x \\ &= \operatorname*{argmin}_{x\in\mathscr{X}\cap\mathscr{D}} \eta g_t^{\top} x + D_{\Phi}\left(x, x_t\right) \end{aligned}.$$

Convergence analysis

Theorem 5.3

Suppose f is convex and Lipschitz continuous (i.e. $\|\boldsymbol{g}^t\|_* \leq L_f$) on $\mathcal{C}$, and suppoe φ is ρ-strongly convex w.r.t. $\|\cdot\|$. Then

$$f^{\mathsf{best},t} - f^{\mathsf{opt}} \leq \frac{\sup_{\boldsymbol{x}\in\mathcal{C}} D_{\varphi}(\boldsymbol{x}, \boldsymbol{x}^0) + \frac{L_f^2}{2\rho}\sum_{k=0}^{t}\eta_k^2}{\sum_{k=0}^{t}\eta_k}$$