Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization

Author: Martin Jaggi
Presenter: Zhongxing Peng

Outline

Outline

Problem Formulation

Constrained convex optimization problem

$$\min_{x \in \mathcal{D}} f(x) \tag{1}$$

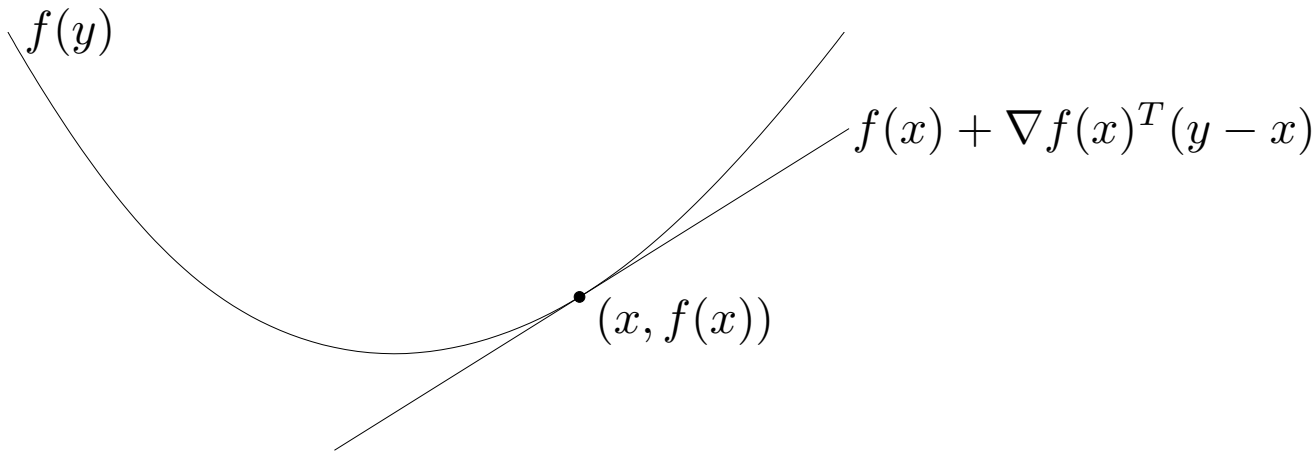
where

1. f is a convex function,
2. $\mathcal{D}$ is a compact convex set.

A set is compact if it is closed and bounded.

Poor Man's Approach

Since the function f is convex, its linear approximation must lie below the graph of the function.



Define dual function $w(x)$ as the minimum of the linear approximation to f at point x over the domain $\mathcal{D}$. Thus, we have weak duality

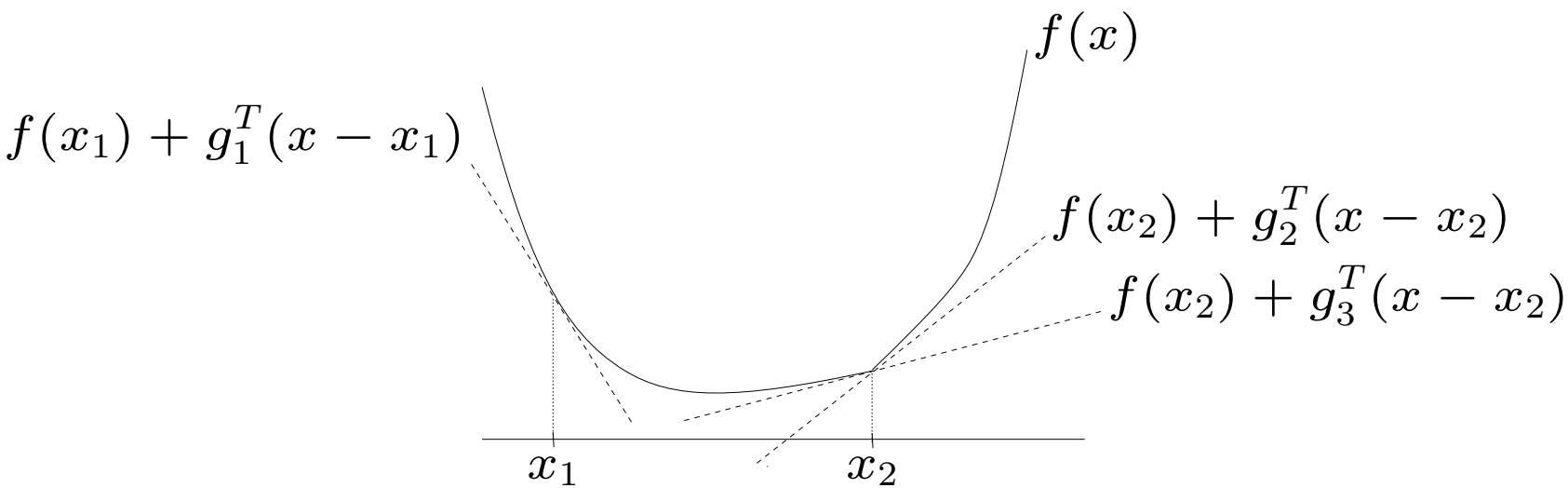
$$w(x) \leq f(y) \tag{2}$$

for each pair $x, y \in \mathcal{D}$.

Subgradient

$d_x \in \partial f(x)$ is called a subgradient to f at x, and belongs to

$$\partial f(x) = \left\{ d_x \in \mathcal{X} \middle| f(y) \geq f(x) + \langle y - x, d_x \rangle, \text{ for } \forall y \in \mathcal{D} \right\} \quad (3)$$



Dual Function

For a given $x \in \mathcal{D}$, and any choice of a subgradient $d_x \in \partial f(x)$, we define a dual function as

$$w(x, d_x) = \min_{y \in \mathcal{D}} f(x) + \langle y - x, d_x \rangle \tag{4}$$

The property of weak-duality is

Lemma 1 (Weak duality)

For all pairs $x, y \in \mathcal{D}$, it holds that

$$w(x, d_x) \leq f(y). \tag{5}$$

If f is differentiable, we have

$$w(x) = w(x, \nabla f(x)) = \min_{\mathcal{D}} f(x) + \langle y - x, \nabla f(x) \rangle. \tag{6}$$

Duality Gap

$g(x, d_x)$ is called the duality gap at x, for the chosen d_x, i.e.

$$g(x, d_x) = f(x) - w(x, d_x) = \max_{y \in \mathcal{D}} \langle x - y, d_x \rangle \tag{7}$$

which is a simple measure of approximation quality.
According to Lemma 1, we have

$$w(x, d_x) \leq f(x^*) \tag{8}$$
$$f(x) - w(x, d_x) \geq f(x) - f(x^*) \tag{9}$$
$$f(x) - w(x, d_x) \geq f(x) - f(x^*) \tag{10}$$
$$g(x, d_x) \geq f(x) - f(x^*) \geq 0. \tag{11}$$

where $f(x) - f(x^*)$ is primal error. If f is differentiable, we have

$$g(x) = g(x, \nabla f(x)) = \max_{y \in \mathcal{D}} \langle x - y, \nabla f(x) \rangle. \tag{12}$$

Relation to Duality of Norms

Observation 1

For optimization over any domain $D = \left\{ x \in \mathcal{X} \middle| \|x\| \leq 1 \right\}$ *being the unit ball of some norm* $\|\cdot\|$*, the duality gap for the optimization problem* $\min_{x \in \mathcal{D}} x$ *is given by*

$$g(x, d_x) = \|d_x\|_* + \langle x, d_x \rangle \tag{13}$$

where $\|\cdot\|_*$ *is the dual norm of* $\|\cdot\|$*.*

Proof.

Since the dual norm is defined as $\|x\|_* = \sup_{\|y\| \leq 1} \langle y, x \rangle$, consider duality gap satisfies

$$g(x, d_x) = \max_{y \in \mathcal{D}} \langle x - y, d_x \rangle = \max_{y \in \mathcal{D}} \langle -y, d_x \rangle + \langle x, d_x \rangle. \tag{14}$$

Frank-Wolfe Algorithm

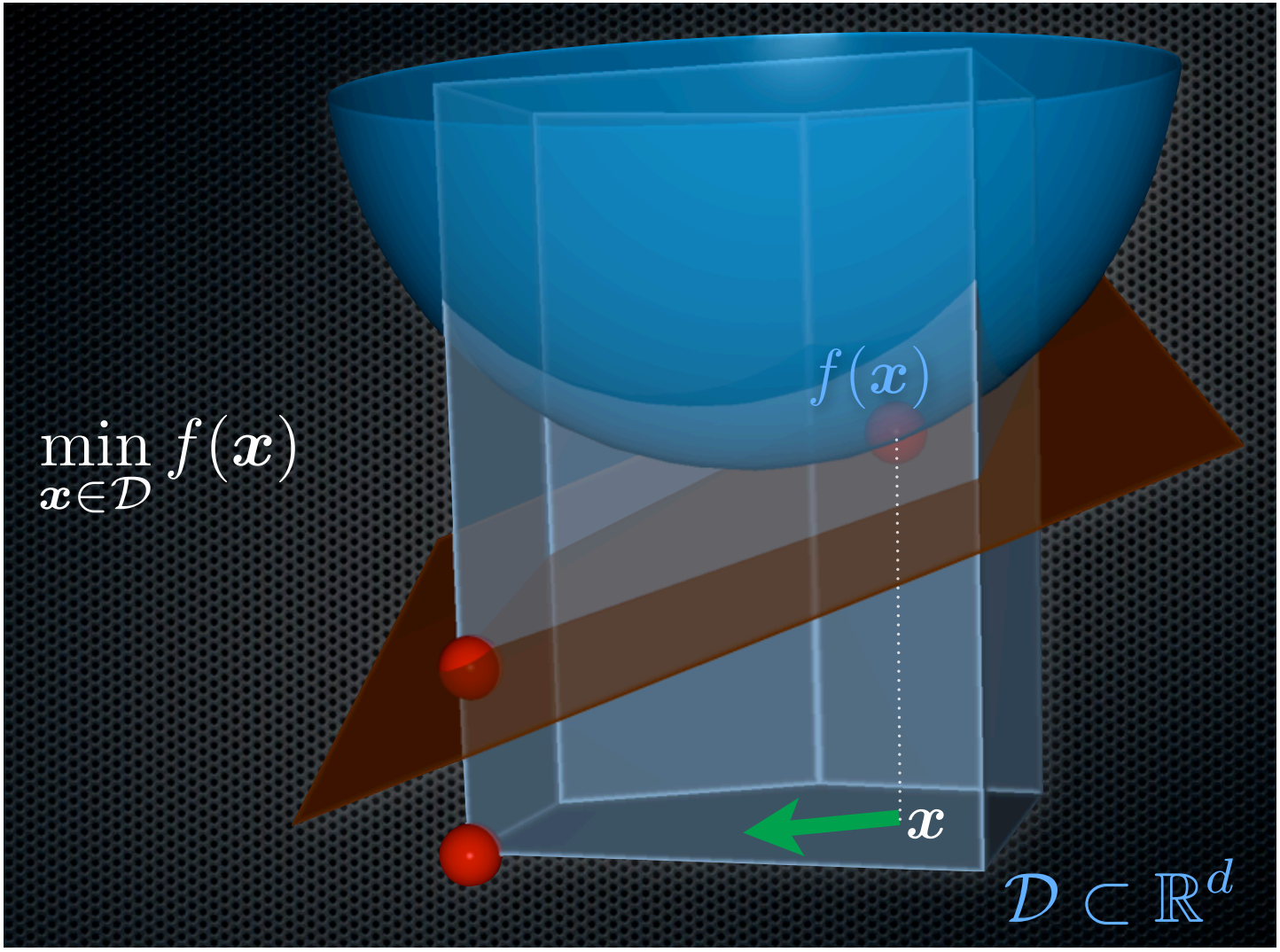
Algorithm 1 Frank-Wolfe (1956)

Let $\boldsymbol{x}^{(0)} \in \mathcal{D}$
for $k = 0 \dots K$ **do**
  Compute $\boldsymbol{s} := \arg\min_{\boldsymbol{s} \in \mathcal{D}} \left\langle \boldsymbol{s}, \nabla f(\boldsymbol{x}^{(k)}) \right\rangle$
  Update $\boldsymbol{x}^{(k+1)} := (1-\gamma)\boldsymbol{x}^{(k)} + \gamma \boldsymbol{s}, \quad \text{for } \gamma := \frac{2}{k+2}$
end for

where we have

$$\begin{aligned} x^{(k+1)} &= (1-\gamma)x^{(k)} + \gamma s \\ &= x^{(k)} + \gamma(s - x^{(k)}) \end{aligned} \tag{15}$$

Geometry Interpretation



Compare to Classical Gradient Descent

Define the descent direction as follows

$$\langle y, \nabla f(x) \rangle < 0 \tag{16}$$

- **Frank Wolfe algorithm:** always choose the best descent direction over the entire domain $\mathcal{D}$
- **Classical gradient descent:**

$$x^{(k+1)} = x^{(k)} + \alpha \nabla f(x^{(k)}) \tag{17}$$

where $\alpha \geq 0$ is the step size.

  - It only uses local information to determine the step-directions.
  - It faces the risk of walking out of the domain $\mathcal{D}$.
  - It requires projection steps after each iteration.

Greedy on a Convex Set

Algorithm 1 Greedy on a Convex Set

Input: Convex function f, convex set D, target accuracy ε
Output: ε-approximate solution for problem (2.1)
Pick an arbitrary starting point $x^{(0)} \in D$
for $k = 0 \ldots \infty$ **do**
  Let $\alpha := \frac{2}{k+2}$
    Compute $s := \textsc{ExactLinear}\left(\nabla f(x^{(k)}), D\right)$
      {Solve the linearized primitive problem exactly}
  —or—
    Compute $s := \textsc{ApproxLinear}\left(\nabla f(x^{(k)}), D, \alpha C_f\right)$
      {Approximate the linearized primitive problem}
  Update $x^{(k+1)} := x^{(k)} + \alpha(s - x^{(k)})$
end for

We call a point $x \in \mathcal{X}$ an ϵ-approximation if $g(x, d_x) \leq \epsilon$ for some choice of subgradient $d_x \in \partial f(x)$.

Linearized Optimization Primitive

In the above algorithm, $\textsc{ExactLinear}(c, \mathcal{D})$ minimizes the linear function $\langle x, c\rangle$ over $\mathcal{D}$. It returns s by

$$s = \arg\min_{y \in \mathcal{D}} \langle y, c\rangle \tag{18}$$

We search for a point s that realizes the current duality gap $g(x)$, that is the distance to the linear approximation, as follows

$$g(x, d_x) = f(x) - w(x, d_x) = \max_{y \in \mathcal{D}} \langle x - y, d_x\rangle. \tag{19}$$

Linearized Optimization Primitive

In the above algorithm, $\textsc{ApproxLinear}(c, \mathcal{D}, \epsilon')$ approximates the minimum of the linear function $\langle x, c \rangle$ over $\mathcal{D}$. It returns s such that

$$\langle s, c \rangle = \arg\min_{y \in \mathcal{D}} \langle y, c \rangle + \epsilon'. \tag{20}$$

For several applications, this can be done significantly more efficiently than the exact variant.

The Curvature

The Curvature constant C_f of a convex and differentiable function f, with respect to a compact domain $\mathcal{D}$ is defined as

$$C_f = \sup_{\substack{x,s\in\mathcal{D} \\ \alpha\in[0,1] \\ y=x+\alpha(s-x)}} \frac{1}{\alpha^2}(f(y) - f(x) - \langle y - x, \nabla f(x)\rangle) \tag{21}$$

- It bounds the gap between $f(y)$ and its linearization.
- $f(y) - f(x) - \langle y - x, \nabla f(x)\rangle$ is known as the Bregman divergence.
- For linear functions f, it holds that $C_f = 0$.

Convergence in Primal Error

Theorem 1 (Primal Convergence)

For each $k \geq 1$, the iterative $x^{(k)}$ of the exact variant of Algorithm 1 satisfies

$$f(x^{(k)}) - f(x^*) \leq \frac{4C_f}{K+2}, \tag{22}$$

where $x^{(*)} \in \mathcal{D}$ is an optimal solution to problem (1). For the approximate variant of Algorithm 1, it holds that

$$f(x^{(k)}) - f(x^*) \leq \frac{8C_f}{K+2}. \tag{23}$$

In other words, both algorithm variants deliver a solution of primal error at most ϵ after $\mathcal{O}(\frac{1}{\epsilon})$ many iterations.

Duality Gap

Theorem 2 (Primal-Dual Convergence)

Let $K = \left\lceil \frac{4C_f}{\epsilon} \right\rceil$. *We run the exact variant of Algorithm 1 for* K *iterations (recall that the step-sizes are given by* $\alpha^{(k)} = \frac{2}{k+2}, 0 \leq k \leq K$*), and then continue for another* $K+1$ *iterations, now with the fixed step-size* $\alpha^{(k)} = \frac{2}{K+2}$ *for* $K \leq k \leq 2K+1$.
Then the algorithm has an iterate $x^{(\hat{k})}$, $K \leq \hat{k} \leq 2K+1$, *with duality gap bounded by*

$$g(x^{(\hat{k})}) \leq \epsilon \tag{24}$$

The same statement holds for the approximate variant of Algorithm 1, when setting $K = \left\lceil \frac{8C_f}{\epsilon} \right\rceil$ *instead.*

Choose Step-Size by Line-Search

Instead of the fixed step-size $\alpha = \frac{2}{k+2}$, we can find the optimal $\alpha \in [0, 1]$ by line-search.

$$\alpha = \underset{\alpha \in [0,1]}{\arg\min} f\left(x^{(k)} + \alpha(s - x^{(k)})\right). \tag{25}$$

If we define

$$f_\alpha = f\left(x_{(\alpha)}^{(k+1)}\right) = f\left(x^{(k)} + \alpha(s - x^{(k)})\right). \tag{26}$$

The optimal α is

$$\frac{\partial}{\partial \alpha} f_\alpha = \left\langle s - x^{(k)}, \nabla f\left(x_{(\alpha)}^{(k+1)}\right)\right\rangle = 0. \tag{27}$$

Greedy on a Convex Set using Line-Search

Algorithm 2 Greedy on a Convex Set, using Line-Search

Input: Convex function f, convex set D, target accuracy ε
Output: ε-approximate solution for problem (3.1)
Pick an arbitrary starting point $x^{(0)} \in D$
for $k = 0 \dots \infty$ **do**
  Compute $s := \textsc{ExactLinear}\left(\nabla f(x^{(k)}), D\right)$
  —or—
  Compute $s := \textsc{ApproxLinear}\left(\nabla f(x^{(k)}), D, \frac{2C_f}{k+2}\right)$
  Find the optimal step-size $\alpha := \arg\min_{\alpha \in [0,1]} f\left(x^{(k)} + \alpha(s - x^{(k)})\right)$
  Update $x^{(k+1)} := x^{(k)} + \alpha(s - x^{(k)})$
end for

Relating the Curvation to the Hessian Matrix

- Hessian matrix of f: is the second derivative of f, i.e. $\nabla^2 f$.
- Second order Taylor-expansion of function f at point x is

$$f(x + \alpha(s - x)) = f(x) + \alpha(s - x)^T \nabla f(x) + \frac{\alpha^2}{2}(s - x)^T \nabla^2 f(z)(s - x) \tag{28}$$

where $z \in [x, y] \subseteq \mathcal{D}$ and

$$y = x + \alpha(s - x) \tag{29}$$

$$\alpha(s - x) = y - x, \tag{30}$$

Thus, we have

$$f(y) = f(x) + (y - x)^T \nabla f(x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x) \tag{31}$$

Relating the Curvation to the Hessian Matrix

$$f(y) - f(x) - \langle y - x, \nabla f(x) \rangle = \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x) \quad (32)$$

According to the definition of C_f

$$C_f = \sup_{\substack{x,s \in \mathcal{D} \\ \alpha \in [0,1] \\ y = x + \alpha(s - x)}} \frac{1}{\alpha^2}(f(y) - f(x) - \langle y - x, \nabla f(x) \rangle), \quad (33)$$

we have

$$C_f \leq \sup_{\substack{x,y \in \mathcal{D} \\ z \in [x,y] \subseteq \mathcal{D}}} \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x) \quad (34)$$

Relating the Curvation to the Hessian Matrix

Lemma 2

For any twice differentiable convex function f over a compact convex domain $\mathcal{D}$, it holds that

$$C_f \leq \frac{1}{2} diam(D)^2 \cdot \sup_{z \in \mathcal{D}} \lambda_{max}(\nabla^2 f(z)) \tag{35}$$

where $diam(\cdot)$ is the Euclidean diameter of the domain.

Relatinig the Curvation to the Hessian Matrix

Proof.

According to Cauchy-Schwarz inequality

$$|\langle a, b\rangle| \leq \|a\| \cdot \|b\| \tag{36}$$

we have

$$\begin{aligned}
(y-x)^T \nabla^2 f(z)(y-x) &\leq \|y-x\|_2 \|\nabla^2 f(z)(y-x)\|_2 && (37)\\
&\leq \|y-x\|_2^2 \frac{\|\nabla^2 f(z)(y-x)\|_2}{\|y-x\|_2} && (38)\\
&\leq \|y-x\|_2^2 \|\nabla^2 f(z)\|_{spec} && (39)\\
&\leq \text{diam}(D)^2 \cdot \sup_{z\in\mathcal{D}} \lambda_{max}(\nabla^2 f(z)) && (40)
\end{aligned}$$

where $\|A\|_{spec} = \sup_{a\neq 0} \frac{\|Aa\|_2}{\|a\|_2}$ is the spectral norm, which is the largest eigenvalue for a positive-semidefinite A. □

VS. Lipschitz-Continuous Gradient

Lemma 3

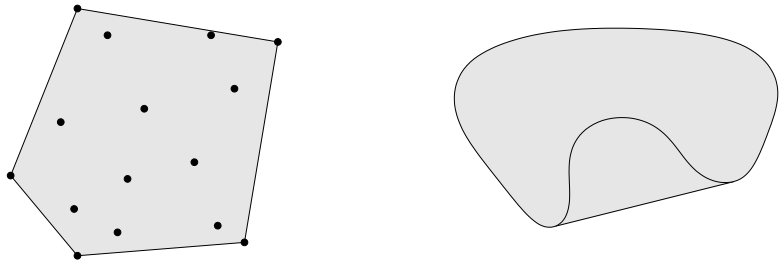
Let f be a convex and twice differential function, and assume that the gradient ∇f is Lipschitz-continuous over the domain $\mathcal{D}$ with Lipschitz-constant $L > 0$. Then

$$C_f \leq \frac{1}{2} diam(\mathcal{D})^2 L \tag{41}$$

where Lipschitz-continuous means there exits $L > 0$ satisfies

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L\|y - x\|_2 \tag{42}$$

Optimizing over Convex Hulls



The convex hull of a set V, denoted $\mathbf{conv}(V)$, is the set of all convex combination of points in V:

$$\mathbf{conv}(V) = \left\{ \theta_1 x_1 + \cdots + \theta_k x_k \middle| x_i \in V, \theta_i \geq 0, i = 1, \cdots, k, \right.$$
$$\left. \theta_1 + \cdots + \theta_k = 1 \right\}. \tag{43}$$

- The convex hull $\mathbf{conv}(V)$ is always convex.
- It is the smallest convex set that contains V.

Optimizing over Convex Hulls

- Consider the case where domain $\mathcal{D}$ is the convex hull of a set V, i.e. $D = \mathbf{conv}(V)$.

Lemma 4 (Linear Optimization over Convex Hulls)

Let $D = \mathbf{conv}(V)$ for any subset $V \subset \mathcal{X}$, and $\mathcal{D}$ compact. Then any linear function $y \mapsto \langle y, c \rangle$ will attain its minimum and maximum over $\mathcal{D}$ at some "vertex" $v \in V$.

- In many applications, the set V is often much easier to describe than the full compact domain D, the result in the above lemma will be usefull to solve the linearized subproblem $\textsc{ExactLinear}()$ more efficiently.

Outline

Sparse Approximation (SA) over the Simplex

The unit Simplex is defined as

$$\Delta_n = \{x \in \mathbb{R}^n | x \geq 0, \|x\|_1 = 1\} \tag{44}$$

Then, optimization on Simplex is

$$\min_{x \in \Delta_n} f(x) \tag{45}$$

Algorithm 3 Sparse Greedy on the Simplex

Input: Convex function f, target accuracy ε
Output: ε-approximate solution for problem (3.1)
Set $x^{(0)} := \mathbf{e}_1$
for $k = 0 \ldots \infty$ **do**
  Compute $i := \arg\min_i \left(\nabla f(x^{(k)})\right)_i$
  Let $\alpha := \frac{2}{k+2}$
  Update $x^{(k+1)} := x^{(k)} + \alpha(\mathbf{e}_i - x^{(k)})$
end for

SA over the Simplex

Theorem 3 (Convergence of Sparse Greedy on the Simplex)

For each $k \geq 1$, the iterate $x^{(k)}$ of the above algorithm satisfies

$$f(x^k) - f(x^*) \leq \frac{4C_f}{k+2} \tag{46}$$

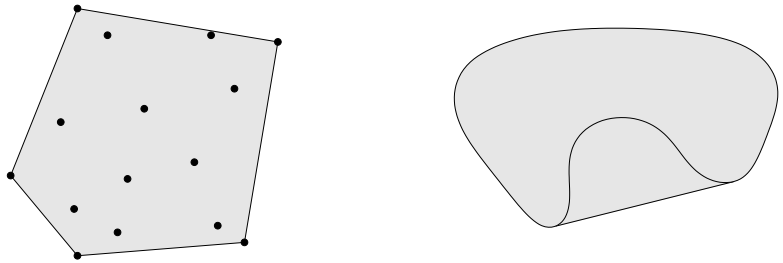
where $x^* \in \Delta_n$ is an optimal solution to problem in (45). Furthermore, for any $\epsilon > 0$, after at most $2\left\lceil \frac{4C_f}{\epsilon} \right\rceil + 1 = \mathcal{O}(\frac{1}{\epsilon})$ many steps, it has an iterate $x^{(k)}$ of sparsity $\mathcal{O}(\frac{1}{\epsilon})$, satisfying $g(x^{(k)}) \leq \epsilon$.

SA over the Simplex

Its duality gap is

$$
\begin{aligned}
g(x, d_x) &= f(x) - w(x, d_x) && (47)\\
&= f(x) - \left(\min_{y \in \mathcal{D}} f(x) + \langle y - x, d_x \rangle \right) && (48)\\
&= \max_{y \in \mathcal{D}} \langle x - y, d_x \rangle && (49)\\
&= x^T d_x - \min_{y \in \mathcal{D}} y^T d_x && (50)
\end{aligned}
$$

SA over the Simplex



The convex hull of a set V, denoted $\mathbf{conv}(V)$, is the set of all convex combination of points in V:

$$\mathbf{conv}(V) = \left\{ \theta_1 x_1 + \cdots + \theta_k x_k \middle| x_i \in V, \theta_i \geq 0, i = 1, \cdots, k, \right.$$
$$\left. \theta_1 + \cdots + \theta_k = 1 \right\}. \tag{51}$$

- The convex hull $\mathbf{conv}(V)$ is always convex.
- It is the smallest convex set that contains V.

SA over the Simplex

Lemma 5 (Linear Optimization over Convex Hulls)

Let $D = \mathbf{conv}(V)$ for any subset $V \subset \mathcal{X}$, and $\mathcal{D}$ compact. Then any linear function $y \mapsto \langle y, c \rangle$ will attain its minimum and maximum over $\mathcal{D}$ at some "vertex" $v \in V$.

Proof.

1. Assume $s \in \mathcal{D}$ satisfies $\langle s, c \rangle = \max_{y \in \mathcal{D}} \langle y, c \rangle$.
2. Represent $s = \sum_i \alpha_i v_i$, where $\sum_i \alpha_i = 1$.
3. We have

$$\langle s, c \rangle = \left\langle \sum_i \alpha_i v_i, c \right\rangle = \sum_i \alpha_i \langle v_i, c \rangle \tag{52}$$

□

SA over the Simplex

Its duality gap is

$$g(x, d_x) = f(x) - w(x, d_x) \tag{53}$$

$$= f(x) - \left(\min_{y \in \mathcal{D}} f(x) + \langle y - x, d_x \rangle \right) \tag{54}$$

$$= \max_{y \in \mathcal{D}} \langle x - y, d_x \rangle \tag{55}$$

$$= x^T d_x - \min_{y \in \mathcal{D}} y^T d_x \tag{56}$$

Because, here we have

$$\mathcal{D} = \Delta_n = \{x \in \mathbb{R}^n | x \geq 0, \|x\|_1 = 1\} \tag{57}$$

then

$$g(x) = g(x, \nabla f(x)) = x^T \nabla f(x) - \min_i (\nabla f(x))_i \tag{58}$$

SA over the Simplex: Example

Consider the following problem

$$\min_{x \in \Delta_n} f(x) = \|x\|_2^2 = x^T x, \tag{59}$$

whose gradient is $\nabla f(x) = 2x$. Then, we have

$$\begin{aligned} f(y) - f(x) - \langle y - x, \nabla f(x) \rangle &= y^T y - x^T x - 2(y - x)^T x && (60) \\ &= \|y - x\|_2^2 && (61) \\ &= \|x + \alpha(s - x) - x\|_2^2 && (62) \\ &= \alpha^2 \|s - x\|_2^2 && (63) \end{aligned}$$

According to the definition of C_f

$$\begin{aligned} C_f &= \sup_{\substack{x,s \in \mathcal{D} \\ \alpha \in [0,1] \\ y = x + \alpha(s - x)}} \frac{1}{\alpha^2} (f(y) - f(x) - \langle y - x, \nabla f(x) \rangle) && (64) \\ &= \sup_{x,s \in \Delta_n} \|x - s\|_2^2 = \mathbf{diam}(\Delta_n)^2 = 2 && (65) \end{aligned}$$

SA with Bounded ℓ_1-Norm

The ℓ_1-ball is defined as

$$\Diamond_n = \{x \in \mathbb{R}^n \big| \|x\|_1 \leq 1\}. \tag{66}$$

Then, optimization with bounded ℓ_1-norm is

$$\min_{x \in \Diamond_n} f(x) \tag{67}$$

Observation 2

For any vector $c \in \mathbb{R}^n$, it holds that

$$e_i \cdot \mathbf{sgn}(c_i) \in \arg\max_{y \in \Diamond_n} y^T c \tag{68}$$

where $i \in \arg\max_j |c_j|$.

SA with Bounded ℓ_1-Norm

Algorithm 4 Sparse Greedy on the ℓ_1-Ball

Input: Convex function f, target accuracy ε
Output: ε-approximate solution for problem (3.3)
Set $x^{(0)} := \mathbf{0}$
for $k = 0 \ldots \infty$ **do**
  Compute $i := \arg\max_i \left|\left(\nabla f(x^{(k)})\right)_i\right|$,
  and let $s := \mathbf{e}_i \cdot \operatorname{sign}\left(\left(-\nabla f(x^{(k)})\right)_i\right)$
  Let $\alpha := \frac{2}{k+2}$
  Update $x^{(k+1)} := x^{(k)} + \alpha(s - x^{(k)})$
end for

SA with Bounded ℓ_1-Norm

Theorem 4 (Convergence of Sparse Greedy on the ℓ_1-Ball)

For each $k \geq 1$, the iterate $x^{(k)}$ of the above algorithm satisfies

$$f(x^k) - f(x^*) \leq \frac{4C_f}{k+2} \tag{69}$$

where $x^* \in \Diamond_n$ is an optimal solution to problem in (67). Furthermore, for any $\epsilon > 0$, after at most $2\left\lceil \frac{4C_f}{\epsilon} \right\rceil + 1 = \mathcal{O}(\frac{1}{\epsilon})$ many steps, it has an iterate $x^{(k)}$ of sparsity $\mathcal{O}(\frac{1}{\epsilon})$, satisfying $g(x^{(k)}) \leq \epsilon$.

Optimization with Bounded ℓ_∞-Norm

The ℓ_∞-ball is defined as

$$\square_n = \{x \in \mathbb{R}^n \big| \|x\|_\infty \leq 1\}. \tag{70}$$

Then, optimization with bounded ℓ_1-norm is

$$\min_{x \in \square_n} f(x) \tag{71}$$

Observation 3

For any vector $c \in \mathbb{R}^n$, it holds that

$$s^c \in \arg\max_{y \in \square_n} y^T c \tag{72}$$

where $(s^c)_i = \mathbf{sgn}(c_i) \in \{-1, 1\}$.

Optimization with Bounded ℓ_∞-Norm

Algorithm 5 Sparse Greedy on the Cube

Input: Convex function f, target accuracy ε

Output: ε-approximate solution for problem (3.4)

Set $x^{(0)} := \mathbf{0}$

for $k = 0 \ldots \infty$ **do**

Compute the sign-vector $\mathbf{s}$ of $\nabla f(x^{(k)})$, such that

$$\mathbf{s}_i = \operatorname{sign}\left(\left(-\nabla f(x^{(k)})\right)_i\right), \quad i = 1..n$$

Let $\alpha := \frac{2}{k+2}$

Update $x^{(k+1)} := x^{(k)} + \alpha(\mathbf{s} - x^{(k)})$

end for

Optimization with Bounded ℓ_∞-Norm

Theorem 5 (Convergence of Sparse Greedy on the ℓ_1-Ball)

For each $k \geq 1$, the iterate $x^{(k)}$ of the above algorithm satisfies

$$f(x^k) - f(x^*) \leq \frac{4C_f}{k+2} \tag{73}$$

where $x^* \in \square_n$ is an optimal solution to problem in (71). Furthermore, for any $\epsilon > 0$, after at most $2\left\lceil \frac{4C_f}{\epsilon} \right\rceil + 1 = \mathcal{O}(\frac{1}{\epsilon})$ many steps, it has an iterate $x^{(k)}$ with $g(x^{(k)}) \leq \epsilon$.

Semidefinite Optimization (SDO) with Bounded Trace

The set of positive semidefinite (PSD) matrices of unit trace is defined as

$$\mathcal{S} = \{X \in \mathbb{R}^{n \times n} | X \succeq 0, \mathrm{Tr}(X) = 1\}. \tag{74}$$

Then, optimization of PSD with bounded trace is

$$\min_{x \in \mathcal{S}} f(x) \tag{75}$$

SDO with Bounded Trace

Algorithm 6 Hazan's Algorithm / Sparse Greedy for Bounded Trace

Input: Convex function f with curvature C_f, target accuracy ε
Output: ε-approximate solution for problem (3.5)
Set $X^{(0)} := vv^T$ for an arbitrary unit length vector $v \in \mathbb{R}^n$.
for $k = 0 \ldots \infty$ **do**
  Let $\alpha := \frac{2}{k+2}$
  Compute $v := v^{(k)} = \text{ApproxEV}\left(\nabla f(X^{(k)}), \alpha C_f\right)$
  Update $X^{(k+1)} := X^{(k)} + \alpha(vv^T - X^{(k)})$
end for

$\textsc{ApproxEV}(A, \epsilon')$ returns v, which satisfies

$$v^T A v \leq \lambda_{min}(A) + \epsilon' \tag{76}$$

SDO with Bounded Trace

Theorem 6

For each $k \geq 1$, the iterate $X^{(k)}$ of the above algorithm satisfies

$$f(X^k) - f(X^*) \leq \frac{8C_f}{k+2} \tag{77}$$

where $X^* \in \mathcal{S}$ is an optimal solution to problem in (75). Furthermore, for any $\epsilon > 0$, after at most $2\left\lceil \frac{8C_f}{\epsilon} \right\rceil + 1 = \mathcal{O}(\frac{1}{\epsilon})$ many steps, it has an iterate $X^{(k)}$ of rank $\mathcal{O}\left(\frac{1}{\epsilon}\right)$, satisfying $g(x^{(k)}) \leq \epsilon$.

Nuclear Norm Regularization

We consider the following problem

$$\min_{Z \in \mathbb{R}^{m \times n}} f(Z) + \mu \|Z\|_* \tag{78}$$

which is equivalent to

$$\min_{Z \in \mathbb{R}^{m \times n}, \|Z\|_* \leq \frac{t}{2}} f(Z) \tag{79}$$

where $f(Z)$ is any differentiable convex function. $\|\cdot\|_*$ is the nuclear norm (trace norm, Schatten 1-norm, the Ky Fan r-norm) of a matrix

$$\|Z\|_* = \sum_{i=1}^{r} \sigma_i(Z) \tag{80}$$

where σ_i is the i-th largest singular value of Z, r is the rank of Z

Nuclear Norm $\approx$ Low Rank

- Frobenius norm:

$$\|Z\|_F = \sqrt{\langle Z, Z \rangle} = \sqrt{\mathrm{Tr}(Z^T Z)}$$
$$= \left(\sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}^2 \right)^{\frac{1}{2}} = \left(\sum_{i=1}^{r} \sigma_i^2 \right)^{\frac{1}{2}} \tag{81}$$

- Operator norm (induced 2-norm):

$$\|Z\| = \sqrt{\lambda_{max}(Z^T Z)} = \sigma_1(Z) \tag{82}$$

- Nuclear norm:

$$\|Z\|_* = \sum_{i=1}^{r} \sigma_i(Z) \tag{83}$$

Nuclear Norm $\approx$ Low Rank

According to the definitions of

$$\|Z\|_F = \left(\sum_{i=1}^{r} \sigma_i^2\right)^{\frac{1}{2}}, \quad \|Z\| = \sigma_1(Z), \quad \|Z\|_* = \sum_{i=1}^{r} \sigma_i(Z),$$

we have following inequalities

$$\|Z\| \leq \|Z\|_F \leq \|Z\|_* \leq \sqrt{r}\|Z\|_F \leq r\|Z\| \tag{84}$$

Nuclear Norm $\approx$ Low Rank

- $\mathcal{C}$ is a given convex set.
- The convex envelop of a (possible nonconvex) function $f : \mathcal{C} \leftrightarrow \mathbb{R}$ is the largest convex function g such that $g(z) \leq f(z)$ for all $z \in \mathcal{C}$.
- g is the best pointwise approximation to f.
- According to the above inequalities, if $\|Z\| \leq 1$, we have

$$rank(Z) \geq \frac{\|Z\|_*}{\|Z\|} \Longrightarrow rank(Z) \geq \|Z\|_* \tag{85}$$

- Nuclear norm is the tightest convex lower bound of the rank function.

Theorem 7

The convex envelop of $rank(Z)$ *on the set* $\{Z \in \mathbb{R}^{m \times n} : \|Z\| \leq 1\}$ *is the nuclear norm* $\|Z\|_*$.

Corollary 1

Any nuclear norm regularized problem

$$\min_{Z \in \mathbb{R}^{m \times n}, \|Z\|_* \leq \frac{t}{2}} f(Z) \tag{86}$$

is equivalent to a bounded trace convex problem

$$\begin{aligned} \min_{X \in \mathbb{S}^{(m+n)\times(m+n)}} \quad & \hat{f}(X) \\ \text{s.t.} \quad & \operatorname{Tr}(X) = t \\ & X \succeq 0 \end{aligned} \tag{87}$$

where $\hat{f}$ *is defined by* $\hat{f}(X) = f(Z)$ *and*

$$X = \begin{pmatrix} V & Z \\ Z^T & W \end{pmatrix} \tag{88}$$

where $V \in \mathbb{S}^{m \times m}$, $W \in \mathbb{S}^{n \times n}$.

Nuclear Norm Regularization

Algorithm 8 Nuclear Norm Regularized Solver

Input: A convex nuclear norm regularized problem (4.2), target accuracy ε

Output: ε-approximate solution for problem (4.2)

1. Consider the transformed symmetric problem for $\hat{f}$, as given by Corollary 4.4

2. Adjust the function $\hat{f}$ so that it first rescales its argument by t

3. Run Hazan's Algorithm 6 for $\hat{f}(X)$ over the domain $X \in \mathcal{S}$.

$$\min_{X \in \mathbb{S}^{(n)\times(n)}} f(X) \quad \text{s.t.} \quad \operatorname{Tr}(X) = 1, \quad X \succeq 0 \tag{89}$$

$$\min_{X \in \mathbb{S}^{(m+n)\times(m+n)}} \hat{f}(X) \quad \text{s.t.} \quad \operatorname{Tr}(X) = t, \quad X \succeq 0 \tag{90}$$

Nuclear Norm Regularization

Corollary 2

After at most $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ *many iterations, Algorithm 8 obtains a solution that is* ϵ *close to the optimum of (86). The algorithm requires a total of* $\tilde{\mathcal{O}}\left(\frac{N_f}{\epsilon^{1.5}}\right)$ *arithmetic operations (with high probability).*

Thank You!