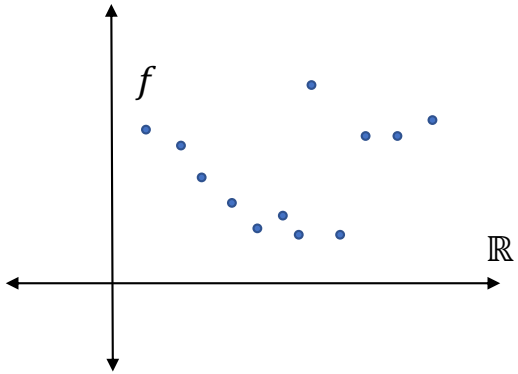
Introduction to Optimization Theory

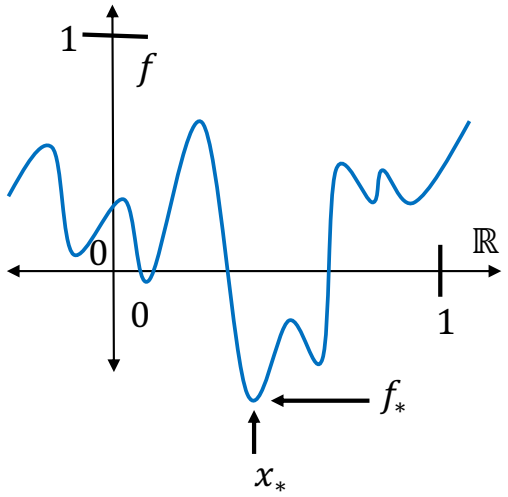
Lecture #16 - 11/10/20

MS&E 213 / CS 2690

Aaron Sidford

sidford@stanford.edu





Plan for Today

Recap

- Online linear optimization methods
- Follow the regularized leader (FTRL)
- Mirror decent

Mirror Descent

- Analysis
- Perspectives
- Comparison

SGD / SVRG

- Finite sum optimization
- Stochastic variance reduced gradient
- Minimizing sum of smooth functions

Dimension Dependent

Cutting plane and interior point methods

Online Linear Optimization

- Given closed convex $S \subseteq \mathbb{R}^n$
- Iterate for round $t \in [T]$
  - We pick $x_t \in S$
  - Adversary reveals penalty $p_t \in \mathbb{R}^n$

Goal:

- Minimize $\text{regret}(T) \stackrel{\text{def}}{=} \sum_{t\in[T]} p_t^\top x_t - \min_{x\in S} \sum_{t\in[T]} p_t^\top x$
- Want average regret $\to 0$: $\lim_{T\to\infty} \frac{\text{regret}(T)}{T} \to 0$
- More generally: bound $\sum_{t\in[T]} p_t^\top (x_t - z)$ for all z

Follow the Regularized Leader (FTRL)

Online Linear Optimization

- For $t \in [T]$
- We pick $x_t \in S$
- Adversary picks $p_t \in \mathbb{R}^n$
- $\sum_{t\in[T]} p_t^\top (x_t - z)$.

Algorithm

- "regularizer" $r: S \to \mathbb{R}$ that is differentiable and μ-strongly convex with respect to $\|\cdot\|$
- Let $\Phi_T(x) \stackrel{\text{def}}{=} \eta \sum_{t\in[T]} p_t^\top x + r(x)$
- Let $\Phi_0(x) \stackrel{\text{def}}{=} r(x)$
- Let $x_{T+1} = \operatorname{argmin}_{x\in S} \Phi_T(x)$

Analysis

- $p_t^\top x_t \le \frac{1}{\eta}[\Phi_t(x_{t+1}) - \Phi_{t-1}(x_t)] + \frac{\eta}{2\mu}\|p_t\|_*^2$
- $\sum_{t\in[T]} p_t^\top (x_t - z) \le \frac{\eta}{2\mu T} \sum_{t\in[T]} \|p_t\|_*^2 + \frac{1}{\eta T}\left[r(z) - \min_{x\in S} r(x)\right]$

Another Online Learning Algorithm

Idea

- Start at x_0
- For $t \in [T]$
- Take small step in "direction" -p_t

How Design Step?

- Pick strongly convex r
- Center regularizer around x_t
  - Notation: $D(x||x_t)$
- Minimizer $\eta \cdot p_t^\top x + D(x||x_t)$

Mirror Descent

Intuition

- *Similar to before*
- *x_t do not move to much relative to reference point*

Centered Regularizer?

Bregman Divergence Distance

- Let r be differentiable and μ-strongly with respect to some norm.
- $D_r(x||c) = r(x) - [r(c) + \nabla r(c)^\top (x - c)]$

Notes

- $D_r(x||c) \geq \frac{\mu}{2} \|x - c\|^2$
- $D_r(x||c) = 0 \Leftrightarrow x = c$
- $D_r(x||y)$ not necessarily $D_r(y||x)$

Examples

- $r(x) = \frac{1}{2} \|x - d\|_2^2$
  - $D_r(x||c) = \frac{1}{2} \|x - c\|_2^2$
  - Doesn't depend on d
- $r(x) = \sum_{i \in [n]} x_i \log x_i$
  - $D_r(x||c) = \sum_{i \in [n]} x_i \log x_i$
  - KL-divergence

Bregman Projection

Bregman Divergence Distance

- Let r be differentiable and μ-strongly with respect to some norm.
- $D_r(x||c) = r(x) - [r(c) + \nabla r(c)^\top (x - y)]$

Bregman Projection

- For closed convex $S \subseteq \mathbb{R}^n$ let $\pi_S^r(y) = \underset{x \in S}{\operatorname{argmin}} D_r(x||y)$

Bregman Pythagorean Theorem (*obtuse angles*)

- $D_r(x||y) + D_r(y||z) \leq D_r(x||z)$ for all $x \in S$
- $\Leftrightarrow y = \pi_S^r(z)$

Proof

- Algebra and optimality of projection

Mirror Descent Analysis

- $D_r(x||c) = r(x) - [r(c) + \nabla r(c)^\top (x - y)]$
- $x_{t+1} = \underset{x \in S}{\operatorname{argmin}}\, \eta \cdot p_t^\top x + D_r(x||x_t)$

Lemmas

- $p_t^\top (x_t - z) \leq \frac{1}{\eta}[D_r(z||x_t) - D_r(z||x_{t+1})] + \frac{\eta}{2\mu}\|p_t\|_*^2$ (*Mirror Descent*)
- $p_t^\top x_t \leq \frac{1}{\eta}[\Phi_t(x_{t+1}) - \Phi_{t-1}(x_t)] + \frac{\eta}{2\mu}\|p_t\|_*^2$ (*FTRL*)

Theorem: Mirror Descent

- $\sum_{t\in[T]} p_t^\top (x_t - z) \leq \sum_{t\in[T]} \frac{\eta}{2\mu}\|p_t\|_*^2 + \frac{1}{\eta} D_r(z||x_0)$

- *Similar bound!*
- *Resulting methods are often the same!*

Theorem: Dual Averaging (FTRL)

- $\sum_{t\in[T]} p_t^\top (x_t - z) \leq \sum_{t\in[T]} \frac{\eta}{2\mu}\|p_t\|_*^2 + \frac{1}{\eta}\left[r(z) - \min_{x\in S} r(x)\right]$

Plan for Today

- Online linear optimization methods
- Follow the regularized leader (FTRL)
- Mirror decent

Mirror Descent

- Analysis
- Perspectives
- Comparison

SGD / SVRG

- Finite sum optimization
- Stochastic variance reduced gradient
- Minimizing sum of smooth functions

Dimension Dependent

Cutting plane and interior point methods

Mirror Descent Analysis

- $D_r(x||c) = r(x) - [r(c) + \nabla r(c)^\top (x - y)]$
- $x_{t+1} = \underset{x \in S}{\operatorname{argmin}}\, \eta \cdot p_t^\top x + D_r(x||x_t)$

Lemma

- $p_t^\top (x_t - z) \leq \frac{1}{\eta} [D_r(z||x_t) - D_r(z||x_{t+1})] + \frac{\eta^2}{2\mu} \|p_t\|_*^2$ (*Mirror Descent*)

Proof

yields "Pythagorean Theorem"

- $D_r(x||z) - D_r(x||y) - D_r(y||z) = \nabla D_r(y||z)^\top (x - y)$ for all x, y, z
- $(\eta \cdot p_t + \nabla D_r(x_{t+1}||x_t))^\top (z - x_{t+1}) \geq 0$
- $\eta \cdot p_t^\top (x_t - z) \leq \eta \cdot p_t^\top (x_t - x_{t+1}) - D_r(x_{t+1}||x_t)$
  $+ D_r(z||x_t) - D_r(z||x_{t+1})$
- $D_r(x_{t+1}||x_t) \geq \frac{\mu}{2} \|x_{t+1} - x_t\|^2$

Another Perspective

- Gradient's are in "dual space"
- Want to map dual to primal to take step

Mirror Map

- Let $r: T \to \mathbb{R}$ be differentiable function with $S \subseteq T$ such that for all $g \in \mathbb{R}^n$ it is the case that $\exists y \in T$ with $\nabla r(y) = g$

Idea

- $x_{t+1} = \pi_S^r(y_t) = \underset{x \in S}{\operatorname{argmin}} D_r(x||y_t)$
- Update $\nabla r(y_t)$ using p_t

Two Algorithms

$$\pi_S^r(y) = \operatorname*{argmin}_{x \in S} D_r(x||y)$$

Algorithm #1

- $y_0 : \nabla r(y_0) = 0$
- For $t \geq 1$
  - $x_t = \pi_S^r(y_t)$
  - $y_{t+1} : \nabla r(y_{t+1}) \coloneqq \nabla r(y_t) - \eta p_t$

FTRL / Dual Averaging

- $\Phi_T(x) \stackrel{\text{def}}{=} \eta \sum_{t \in [T]} p_t^\top x + r(x)$
- Let $x_{T+1} = \operatorname{argmin}_{x \in S} \Phi_T(x)$

- $y_t : \nabla r(y_t) = -\eta \sum_{t \in [T]} p_t^\top x$

Algorithm #2

- $y_0 = x_1$
- For $t \geq 1$
  - $x_t = \pi_S^r(y_t)$
  - $y_{t+1} : \nabla r(y_{t+1}) \coloneqq \nabla r(x_t) - \eta p_t$

Mirror Descent

- $x_{t+1} = \operatorname*{argmin}_{x \in S} \eta \cdot p_t^\top x + D_r(x||x_t)$

- $y_t : \nabla r(y_t) = -\eta p_t + \nabla r(x_{t-1})$

Plan for Today

- Online linear optimization methods
- Follow the regularized leader (FTRL)
- Mirror decent

- Analysis
- Perspectives
- Comparison

- Finite sum optimization
- Stochastic variance reduced gradient
- Minimizing sum of smooth functions

Cutting plane and interior point methods

Motivating Problem

Finite Sum Optimization

- $\min_{x \in \mathbb{R}^n} F(x) = \frac{1}{m} \sum_{i \in [m]} f_i(x)$
- Each f_i is convex and L-smooth
- F is μ-strongly convex
- Have gradient oracle for f_i and want ϵ-optimal point

Note

- Natural machine learning problem
  - x = model, $i =$ data point, $f_i =$ loss of model on data point i
- Many variants

Baseline?

- Gradient Descent
  - F is at most $m \cdot \frac{1}{m} \cdot L$-smooth
  - $O(m(L/\mu) \log \epsilon^{-1})$ queries
- AGD solves
  - $O(m\sqrt{L/\mu} \log \epsilon^{-1})$ queries
- Can we improve?

Example

- $\min_{x\in\mathbb{R}^n} F(x) = \frac{1}{m}\sum_{i\in[m]} f_i(x)$
- Each f_i is convex and L-smooth
- F is μ-strongly convex
- Have gradient oracle for f_i and want ϵ-optimal point

Example Problem: Regression

- $\min_{x\in\mathbb{R}^n} F(x) = \frac{1}{2}\|\boldsymbol{A}x - b\|_2^2$
- $\boldsymbol{A}$ has rows $a_1, \dots, a_m \in \mathbb{R}^n$
- $f_i(x) = \frac{m}{2}\left(a_i^\top x - b\right)^2$
- $F(x) = \frac{1}{m}\sum_{i\in[m]} f_i(x)$
- Each f_i is $\|a_i\|_2^2 m \le L$ smooth

Baseline

- F is $\mu = \lambda_{\min}(\boldsymbol{A}^\top\boldsymbol{A})$-strongly convex
- F is $\lambda_{\max}(\boldsymbol{A}^\top\boldsymbol{A})$-smooth
- $\lambda_{\max}(\boldsymbol{A}^\top\boldsymbol{A}) \le \mathrm{tr}(\boldsymbol{A}^\top\boldsymbol{A}) = \sum_{i\in[m]}\|a_i\|_2^2 \le L$
  - Equal in the worst case
  - Can be better
    - $\lambda_{\max}(\boldsymbol{A}^\top\boldsymbol{A}) \ge \mathrm{tr}(\boldsymbol{A}^\top\boldsymbol{A})/d$
- GD: $O(m(L/\mu)\log\epsilon^{-1})$ rows
- AGD: $O(m\sqrt{L/\mu}\log\epsilon^{-1})$ rows
- Can we improve?

Stochastic Gradient Descent (SGD)

Problem

- Differentiable $f: \mathbb{R}^n \to \mathbb{R}$
- Goal: $\min_x f(x)$

Algorithm

- For $t \geq 1$
  - g_t random s.t. $\mathbb{E}g_t = \nabla f(x_t)$
  - $x_{t+1} = x_t + \eta_t g_t$

Notes

- Many many variant
- Can have $g_t \in \partial f(x_t)$
- Stochastic mirror descent
- Various η_t and x_t aggregation schemes
- Etc.
- **Today**: simple case of differentiable convex f

SGD

- $\min_{x \in \mathbb{R}^n} F(x) = \frac{1}{m} \sum_{i \in [m]} f_i(x)$
- Each f_i is convex and L-smooth
- F is μ-strongly convex
- Have gradient oracle for f_i and want ϵ-optimal point
- x_* minimizes and $\epsilon_t \stackrel{\text{def}}{=} \mathbb{E} f(x_t) - f_*$

SGD Step: If $x_{t+1} = x_t - \eta g_t$ and $\mathbb{E} g_t = \nabla f(x_t)$

- $\epsilon_t \leq \frac{1}{2\eta} [\|x_t - x_*\|_2^2 - \mathbb{E}\|x_{t+1} - x_*\|_2^2] + \frac{\eta}{2} \mathbb{E}\|g_t\|_2^2$
- $\frac{1}{T} \sum_{t \in [T]} \epsilon_t \leq \frac{1}{T\eta\mu} \cdot \epsilon_1 + \frac{\eta}{2T} \sum_{t \in [T]} \mathbb{E}\|g_t\|_2^2$

Question
How pick g_t ?

Proof:

- $\|x_{t+1} - x_*\|_2^2 = \|x_t - x_* - \eta g_t\|_2^2 = \|x_t - x_*\|_2^2 - 2\eta g_t^\top (x_t - x_*) + \eta^2 \|g_t\|_2^2$
- $\mathbb{E} g_t^\top (x_t - x_*) = \nabla f(x_t)^\top (x_t - x_*) = -\nabla f(x_t)^\top (x_* - x_t) \leq -[f(x_t) - f_*]$
- $\|x_{T+1} - x_*\|_2^2 \geq 0$
- $\|x_1 - x_*\|_2^2 \leq \frac{2}{\mu} \cdot \epsilon_1$

Note: will be a little informal (e.g. with expectations in slides, will be more formal in notes.

Random Function

- $\min_{x \in \mathbb{R}^n} F(x) = \frac{1}{m} \sum_{i \in [m]} f_i(x)$
- Each f_i is convex and L-smooth
- F is μ-strongly convex
- Have gradient oracle for f_i and want ϵ-optimal point
- x_* minimizes and $\epsilon_t \stackrel{\text{def}}{=} \mathbb{E} f(x_t) - f_*$

SGD Step: If $x_{t+1} = x_t + \eta g_t$ and $\mathbb{E} g_t = \nabla f(x_t)$

- $\frac{1}{T} \sum_{t \in [T]} \epsilon_t \leq \frac{1}{T\eta\mu} \cdot \epsilon_1 + \frac{\eta}{2T} \sum_{t \in [T]} \mathbb{E} \|g_t\|_2^2$
- **Idea**: $g_t = \nabla f_{i_t}(x_t)$ for random $i_t \in [m]$

Problem

Can be non-zero and large!

Note

Are prominent cases where is 0, e.g. $\epsilon_i = 0$

(Kaczmarz) $Ax = b$ where $\exists x_*$ s.t. $Ax_* = b$

Example

$f_i(x) = \left(a_i^\top x - b_i\right)^2$

and $b_i = a_i^\top x_* + \epsilon_i$ where $\mathbb{E}\epsilon_i = 0$ and $m \to \infty$

Analysis

- $\mathbb{E} \|g_t\|_2^2 = \mathbb{E}_{i_k} \left\| \nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_*) + \nabla f_{i_k}(x_*) \right\|_2^2$
- $\mathbb{E} \|g_t\|_2^2 \leq 2 \cdot \mathbb{E}_{i_k} \left\| \nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_*) \right\|_2^2 + 2 \cdot \mathbb{E}_{i_k} \left\| \nabla f_{i_k}(x_*) \right\|_2^2$
- $\left\| \nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_*) \right\|_2^2 \leq 2L \cdot \left[f_{i_k}(x_k) - \left[f_{i_k}(x_*) - \nabla f_{i_k}(x_*)^\top (x_k - x_*) \right] \right]$
- $\mathbb{E}_{i_k} \left\| \nabla f_{i_k}(x_k) - \nabla f_{i_k}(x_*) \right\|_2^2 \leq 2L \cdot [F(x_k) - F_*] \leq 2L \cdot \epsilon_k$

Great!

Relative error

Decreasing "Variance"

- $\min_{x \in \mathbb{R}^n} F(x) = \frac{1}{m} \sum_{i \in [m]} f_i(x)$
- Each f_i is convex and L-smooth
- F is μ-strongly convex
- Have gradient oracle for f_i and want ϵ-optimal point
- x_* minimizes and $\epsilon_t \stackrel{\text{def}}{=} \mathbb{E} f(x_t) - f_*$

Problem

- Want $\mathbb{E}\|g_t\|_2^2$ small relative to the ϵ_i
- Difficulty: even $\mathbb{E}_{i_k} \left\| \nabla f_{i_k}(x_*) \right\|_2^2 \neq 0$

Idea: add mean $\vec{0}$ vector to improve variance

- $g_t = \nabla f_{i_t}(x_t) - \nabla f_{i_t}(x_*)$ **Pro**: $\mathbb{E}\|g_t\|_2^2 = 0$ when $x_1 = x_*$ **Con**: need know x_*
- $g_t = \nabla f_{i_t}(x_t) - \nabla f_{i_t}(x_1)$ **Pro**: $\mathbb{E}\|g_t\|_2^2 = 0$ when $x_1 = x_*$ **Con**: biased!
- $g_t = \nabla f_{i_t}(x_t) - \nabla f_{i_t}(x_1) + \nabla F(x_1)$

m-queries for initial g_1 then 1 query for each additional g_t

Stochastic Variance Reduced Gradient (SVRG)
Johnson Zhang 13

SVRG Variance

- $\min_{x\in\mathbb{R}^n} F(x) = \frac{1}{m}\sum_{i\in[m]} f_i(x)$
- Each f_i is convex and L-smooth
- F is μ-strongly convex
- Have gradient oracle for f_i and want ϵ-optimal point
- x_* minimizes and $\epsilon_t \stackrel{\text{def}}{=} \mathbb{E}f(x_t) - f_*$

SVRG Step: $g_t = \nabla f_{i_t}(x_t) - \nabla f_{i_t}(x_1) + \nabla F(x_1)$ for random $i_t \in [m]$

- $\mathbb{E}\|g_t\|_2^2 = \mathbb{E}\left\|\nabla f_{i_t}(x_t) - \nabla f_{i_t}(x_*) - \left[\nabla f_{i_t}(x_1) - \nabla f_{i_t}(x_*)\right] + \nabla F(x_1)\right\|_2^2$
- $\mathbb{E}\|g_t\|_2^2 \le 2\cdot\mathbb{E}\left\|\nabla f_{i_t}(x_t) - \nabla f_{i_t}(x_*)\right\|_2^2$
  $+2\cdot\mathbb{E}\left\|\nabla f_{i_t}(x_1) - \nabla f_{i_t}(x_*) - \nabla F(x_1)\right\|_2^2$
- $\mathbb{E}\|v - \mathbb{E}v\|_2^2 \le \mathbb{E}\|v\|_2^2 \Rightarrow$
- $\mathbb{E}\|g_t\|_2^2 \le 2\cdot\left[\mathbb{E}\left\|\nabla f_{i_t}(x_t) - \nabla f_{i_t}(x_*)\right\|_2^2 + \mathbb{E}\left\|\nabla f_{i_t}(x_1) - \nabla f_{i_t}(x_*)\right\|_2^2\right]$
- $\mathbb{E}\|g_t\|_2^2 \le 4L[\epsilon_t + \epsilon_1]$

SVRG

- $\min_{x \in \mathbb{R}^n} F(x) = \frac{1}{m} \sum_{i \in [m]} f_i(x)$
- Each f_i is convex and L-smooth
- F is μ-strongly convex
- Have gradient oracle for f_i and want ϵ-optimal point
- x_* minimizes and $\epsilon_t \stackrel{\text{def}}{=} \mathbb{E} f(x_t) - f_*$

SVRG: $x_{t+1} = x_t + \eta g_t$ and $g_t = \nabla f_{i_t}(x_t) - \nabla f_{i_t}(x_1) + \nabla F(x_1)$ for i_t uniform

- $\Rightarrow \frac{1}{T} \sum_{t \in [T]} \epsilon_t \leq \frac{1}{T \eta \mu} \cdot \epsilon_1 + \frac{\eta}{2T} \sum_{t \in [T]} \mathbb{E} \|g_t\|_2^2$
- $\Rightarrow \frac{1}{T} \sum_{t \in [T]} \epsilon_t \leq \frac{1}{T} \left[\frac{1}{\eta \mu} \cdot \epsilon_1 + \frac{\eta}{2} \cdot 4L \sum_{t \in [T]} [\epsilon_t + \epsilon_1] \right]$
- $\Rightarrow \frac{1-2\eta L}{T} \sum_{t \in [T]} \epsilon_t \leq \frac{1}{\eta \mu T} \cdot \epsilon_1 + 2L\eta \epsilon_1$
- $\left(\eta = \frac{1}{8L} \right) \Rightarrow \frac{1}{T} \sum_{t \in [T]} \epsilon_T \leq \frac{16L}{\mu T} \cdot \epsilon_1 + \frac{1}{2} \epsilon_1$
- $\left(T = \frac{64L}{\mu} \right) \Rightarrow \frac{1}{T} \sum_{t \in [T]} \epsilon_T \leq \frac{3}{4} \epsilon_1$

Query Complexity?

- Can halve error in expectation with $m + T = O(m + L/\mu)$ queries.
- Repeating yields expected ϵ-optimal in $O((m + L/\mu) \log \epsilon^{-1})$ queries!
- GD: $O(m \cdot (L/\mu) \cdot \log \epsilon^{-1})$

State-of-the-art?

- $\min_{x \in \mathbb{R}^n} F(x) = \frac{1}{m} \sum_{i \in [m]} f_i(x)$
- Each f_i is convex and L-smooth
- F is μ-strongly convex
- Have gradient oracle for f_i and want ϵ-optimal point
- x_* minimizes and $\epsilon_t \stackrel{\text{def}}{=} \mathbb{E} f(x_t) - f_*$

- SVRG: $O((m + L/\mu) \log \epsilon^{-1})$ queries
  - Are other ways to achieve
- GD: $O(m \cdot (L/\mu) \cdot \log \epsilon^{-1})$ queres
- AGD: $O(m \cdot (L/\mu) \cdot \log \epsilon^{-1})$ queres
- Accelerated proximal point / Catalyst
  - $\tilde{O}\left(m + \sqrt{mL/\mu}\right) \log \epsilon^{-1})$
- Katyusha
  - $O\left(m + \sqrt{mL/\mu}\right) \log \epsilon^{-1})$

- Many generalization!
- Many more applications!
  - MDPs [**S**WWY18, **S**WWYY18]
  - Submodular [CL**S**W17, HRRS19, AL**S**19]
  - Minimax [CJ**S**T19]
  - Regression, non-convex, …

Plan for Today

- Online linear optimization methods
- Follow the regularized leader (FTRL)
- Mirror decent

- Analysis
- Perspectives
- Comparison

- Finite sum optimization
- Stochastic variance reduced gradient
- Minimizing sum of smooth functions

Cutting plane and interior point methods