Nesterov’s
Optimal Gradient Methods

Xinhua Zhang
Australian National University
NICTA
Outline

- The problem from machine learning perspective
- Preliminaries
  - Convex analysis and gradient descent
- Nesterov’s optimal gradient method
  - Lower bound of optimization
  - Optimal gradient method
- Utilizing structure: composite optimization
  - Smooth minimization
  - Excessive gap minimization
- Conclusion
Outline

- The problem from machine learning perspective
- Preliminaries
  - Convex analysis and gradient descent
- Nesterov’s optimal gradient method
  - Lower bound of optimization
  - Optimal gradient method
- Utilizing structure: composite optimization
  - Smooth minimization
  - Excessive gap minimization
- Conclusion
Many machine learning problems have the form

\[ \min_w J(w) := \lambda \Omega(w) + R_{\text{emp}}(w) \]

where

\[ R_{\text{emp}}(w) := \frac{1}{n} \sum_{i=1}^{n} l(x_i, y_i; w) \]

\[ w : \text{weight vector} \]

\[ \{x_i, y_i\}_{i=1}^{n} : \text{training data} \]

\[ l(x, y; w) : \text{convex and non-negative loss function} \]

\[ \Omega(w) : \text{convex and non-negative regularizer} \]

Can be non-smooth, possibly non-convex.
The problem: Examples

\[
\min_w \frac{\lambda}{2} \|w\|^2 + \frac{1}{n} \sum_{i=1}^n \xi_i \\
\text{s.t. } \xi_i \geq 1 - y_i \langle w, x_i \rangle \quad \forall 1 \leq i \leq n \\
\xi_i \geq 0 \quad \forall 1 \leq i \leq n
\]

\[
\xi_i = \max \{0, 1 - y_i \langle w, x_i \rangle\} \\
\frac{\lambda}{2} \|w\|^2 + \frac{1}{n} \sum_{i=1}^n \max \{0, 1 - y_i \langle w, x_i \rangle\}
\]

<table>
<thead>
<tr>
<th>Model (obj)</th>
<th>$\lambda \Omega(w)$</th>
<th>$R_{\text{emp}}(w)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>linear SVMs</td>
<td>$\frac{\lambda}{2} |w|^2$</td>
<td>$\frac{1}{n} \sum_{i=1}^n \max {0, 1 - y_i \langle w, x_i \rangle}$</td>
</tr>
<tr>
<td>$\ell_1$ logistic regression</td>
<td>$\lambda |w|_1$</td>
<td>$\frac{1}{n} \sum_{i=1}^n \log (1 + \exp (-y_i \langle w, x_i \rangle))$</td>
</tr>
<tr>
<td>$\epsilon$-insensitive classify</td>
<td>$\frac{\lambda}{2} |w|^2$</td>
<td>$\frac{1}{n} \sum_{i=1}^n \max {0,</td>
</tr>
</tbody>
</table>

\[
\|w_1\|_1 = \sum_i |w_i|
\]
The problem:
More examples

Lasso
\[
\arg\min_w \lambda \cdot \|w\|_1 + \|Aw - b\|_2^2
\]

Multi-task learning
\[
\arg\min_w \lambda \cdot \|W\|_{tr} + \sum_{t=1}^{T} \|X_t w_t - b_t\|_2^2
\]
\[
\arg\min_w \lambda \cdot \|W\|_{1,\infty} + \sum_{t=1}^{T} \|X_t w_t - b_t\|_2^2
\]

Matrix game
\[
\arg\min_{w\in\Delta_d} \langle c, w \rangle + \max_{u\in\Delta_n} \{ \langle Aw, u \rangle + \langle b, u \rangle \}
\]

Entropy regularized LPBoost
\[
\arg\min_{w\in\Delta_d} \lambda \Delta(w, w^0) + \max_{u\in\Delta_n} \langle Aw, u \rangle
\]
The problem:

Lagrange dual

Binary SVM

$$\min \frac{1}{2\lambda} \alpha^\top Q \alpha - \sum_i \alpha_i$$

s.t.  \( \alpha_i \in [0, n^{-1}] \)

$$\sum_i y_i \alpha_i = 0$$

Entropy regularized LPBoost

$$\lambda \ln \sum_d w_d^0 \exp \left( -\lambda^{-1} \left( \sum_{i=1}^n A_{i,d} \alpha_i \right) \right)$$

s.t.  \( \alpha_i \in [0, 1] \)

$$\sum_i \alpha_i = 1$$

where

$$Q_{ij} = y_i y_j \langle x_i, x_j \rangle$$
The problem

Summary

\[
\min_{w \in Q} J(w)
\]

where

- \( J \) is convex, but might be non-smooth
- \( Q \) is a (simple) convex set
- \( J \) might have composite form

Solver: iterative method \( w_0, w_1, w_2, \ldots \)

- Want \( \epsilon_k := J(w_k) - J(w^*) \) to decrease to 0 quickly

where \( w^* := \arg\min_{w \in Q} J(w) \).

We only discuss optimization in this session, no generalization bound.
The problem: What makes a good optimizer?

- Find an $\epsilon$-approximate solution $w_k$
  \[ J(w_k) \leq \min_{w} J(w) + \epsilon \]

- Desirable:
  - $k$ as small as possible (take as few steps as possible)
    - Error $\epsilon_k$ decays by $1/k^2$, $1/k$, or $e^{-k}$.
  - Each iteration costs reasonable amount of work
  - Depends on $n$, $\lambda$ and other condition parameters leniently
  - General purpose, parallelizable (low sequential processing)
  - Quit when done (measurable convergence criteria)
The problem:
Rate of convergence

- Convergence rate:

$$\lim_{k \to \infty} \frac{\epsilon_{k+1}}{\epsilon_k} = \begin{cases} 
0 & \text{superlinear rate} \\
\in (0, 1) & \text{linear rate} \\
1 & \text{sublinear rate}
\end{cases}$$

- Use interchangeably:
  - Fix step index $k$, upper bound $\min_{1 \leq t \leq k} \epsilon_t$
  - Fix precision $\epsilon$, how many steps needed for $\min_{1 \leq t \leq k} \epsilon_t < \epsilon$

  - E.g. $\frac{1}{\epsilon^2}, \frac{1}{\epsilon}, \frac{1}{\sqrt{\epsilon}}, \log \frac{1}{\epsilon}, \log \log \frac{1}{\epsilon}$
The problem: Collection of results

- Convergence rate:

<table>
<thead>
<tr>
<th>Objective function</th>
<th>Smooth</th>
<th>Smooth and very convex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient descent</td>
<td>$O\left(\frac{1}{\varepsilon}\right)$</td>
<td>$O\left(\log\frac{1}{\varepsilon}\right)$</td>
</tr>
<tr>
<td>Nesterov</td>
<td>$O\left(\sqrt{\frac{1}{\varepsilon}}\right)$</td>
<td>$O\left(\log\frac{1}{\varepsilon}\right)$</td>
</tr>
<tr>
<td>Lower bound</td>
<td>$O\left(\sqrt{\frac{1}{\varepsilon}}\right)$</td>
<td>$O\left(\log\frac{1}{\varepsilon}\right)$</td>
</tr>
</tbody>
</table>

- Composite non-smooth

Smooth + (dual of smooth)      (very convex) + (dual of smooth)

$O\left(\frac{1}{\varepsilon}\right)$ $O\left(\sqrt{\frac{1}{\varepsilon}}\right)$
Outline

- The problem from machine learning perspective
- Preliminaries
  - Convex analysis and gradient descent
- Nesterov’s optimal gradient method
  - Lower bound of optimization
  - Optimal gradient method
- Utilizing structure: composite optimization
  - Smooth minimization
  - Excessive gap minimization
- Conclusion
A function $f$ is convex iff

$$\forall \mathbf{x}, \mathbf{y}, \lambda \in (0, 1)$$

$$f(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y})$$
Preliminaries: convex analysis

Convex functions

- A function $f$ is convex iff

$$\forall \ x, y, \lambda \in (0, 1)$$

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$$
A function $f$ is called $\sigma$-strongly convex wrt a norm $\| \cdot \|$ iff

$$f(x) - \frac{1}{2}\sigma \|x\|^2$$

is convex

$$\forall x, y, \lambda \in (0, 1)$$

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \sigma \cdot \frac{\lambda(1 - \lambda)}{2} \|x - y\|^2$$
Preliminaries: convex analysis

Strong convexity

- First order equivalent condition

\[ f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\sigma}{2} \|x - y\|^2 \quad \forall \ x, y \]
Preliminaries: convex analysis

Strong convexity

- First order equivalent condition

\[ f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\sigma}{2} \|x - y\|^2 \quad \forall \ x, y \]
Preliminaries: convex analysis

Strong convexity

- Second order

\[ \langle \nabla^2 f(x) y, y \rangle \geq \sigma \|y\|^2 \quad \forall \ x, y \]

- If \( \| \cdot \| \) Euclidean norm, then

\[ \nabla^2 f(x) \succeq \sigma \mathbb{I} \]

- Lower bounds rate of change of gradient
Preliminaries: convex analysis

Lipschitz continuous gradient

- Lipschitz continuity
  - Stronger than continuity, weaker than differentiability
  - Upper bounds rate of change

\[ \exists L > 0 \]

\[ |f(x) - f(y)| \leq L \|x - y\| \quad \forall x, y \]
Preliminaries: convex analysis
Lipschitz continuous gradient

- Gradient is Lipschitz continuous (must be differentiable)

\[ \| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \| \quad \forall x, y \]

\[ f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \| x - y \|^2 \quad \forall x, y \]
Preliminaries: convex analysis

Lipschitz continuous gradient

- Gradient is Lipschitz continuous (must be differentiable)

\[ \| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \| \quad \forall \ x, y \]

\[ f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \| x - y \|^2 \quad \forall \ x, y \]

\[ \langle \nabla^2 f(x)y, y \rangle \leq L \| y \|^2 \quad \forall \ x, y \]

\[ \nabla^2 f(x) \preceq LI \quad \text{if } L_2 \text{ norm} \]
Preliminaries: convex analysis

Fenchel Dual

- Fenchel dual of a function $f$

$$f^*(s) = \sup_{x} \langle s, x \rangle - f(x)$$

- Properties

$$f^{**} = f \quad \text{if } f \text{ is convex and closed}$$

$$\sigma \text{ strongly convex} \quad \longleftrightarrow \quad \frac{1}{\sigma} \text{-l.c.g on } \mathbb{R}^d$$

$$\text{L-l.c.g on } \mathbb{R}^d \quad \longleftrightarrow \quad \frac{1}{L} \text{ strongly convex}$$
Fenchel Dual

Fenchel dual of a function $f$

$$f^*(s) = \sup_x \langle s, x \rangle - f(x)$$

$s = \nabla f(x)$

$s \in \partial f(x)$
Preliminaries: convex analysis: Subgradient

- Generalize gradient to non-differentiable functions
  - Idea: tangent plane lying below the graph of $f$
Preliminaries: convex analysis: Subgradient

- Generalize gradient to non-differentiable functions
  - \( \mu \) is called a subgradient of \( f \) at \( x \) if
    \[
    f(x') \geq f(x) + \langle x' - x, \mu \rangle \quad \forall x'
    \]

- All such \( \mu \) comprise the subdifferential of \( f \) at \( x \): \( \partial f(x) \)
Preliminaries: convex analysis: Subgradient

- Generalize gradient to non-differentiable functions
  - \( \mu \) is called a subgradient of \( f \) at \( x \) if
    \[
    f(x') \geq f(x) + \langle x' - x, \mu \rangle \quad \forall x'
    \]

- All such \( \mu \) comprise the subdifferential of \( f \) at \( x \): \( \partial f(x) \)
- Unique if \( f \) is differentiable at \( x \)
Preliminaries: optimization: Gradient descent

- Gradient descent

\[ x_{k+1} = x_k - s_k \nabla f(x_k), \quad s_k \geq 0 \]

- Suppose \( f \) is both \( \sigma \)-strongly convex and \( L \)-l.c.g.

\[ \epsilon_k := f(x_k) - f(w^*) \]

\[ \epsilon_k \leq \left( 1 - \frac{\sigma}{L} \right)^k \epsilon_0 \]

- Key idea
  - Norm of gradient upper bounds how far away from optimal
  - Lower bounds how much progress one can make
Preliminaries: optimization: Gradient descent

- Upper bound distance from optimal

\[
x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)
\]

\[
slope = \sigma
\]

\[
\begin{align*}
shaded \text{ area} & \leq \text{ triangle area} \\
\frac{1}{2\sigma} \| \nabla f(x_k) \|^2 & = f(x_k) - f(x^*)
\end{align*}
\]

So

\[
f(x_k) - f(x^*) \leq \frac{1}{2\sigma} \| \nabla f(x_k) \|^2
\]
Preliminaries: optimization: Gradient descent

- Lower bound progress at each step

\[
x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)
\]

Shaded area \geq triangle area

\[
f(x_k) - f(x_{k+1}) \geq \frac{1}{2L} \| \nabla f(x_k) \|^2
\]

So

\[
f(x_k) - f(x_{k+1}) \geq \frac{1}{2L} \| \nabla f(x_k) \|^2
\]
Putting things together

$$2\sigma (f(x_k) - f(x^*)) \leq ||\nabla f(x_k)||^2 \leq 2L(f(x_k) - f(x_{k+1}))$$

$$f(x_{k+1}) - f(x^*) \leq (1 - \frac{\sigma}{L}) (f(x_k) - f(x^*))$$
Preliminaries: optimization: Gradient descent

- Putting things together

\[ 2\sigma(f(x_k) - f(x^*)) \leq \|\nabla f(x_k)\|^2 \leq 2L(f(x_k) - f(x_{k+1})) \]

\[ f(x_{k+1}) - f(x^*) \leq (1 - \frac{\sigma}{L}) \epsilon_k \epsilon_{k+1} \]
Preliminaries: optimization: Gradient descent

- Putting things together

\[ 2\sigma(f(x_k) - f(x^*)) \leq \|\nabla f(x_k)\|^2 \leq 2L(f(x_k) - f(x_{k+1})) \]

\[ f(x_{k+1}) - f(x^*) \leq (1 - \frac{\sigma}{L}) (f(x_k) - f(x^*)) \]

- What if \( \sigma = 0 \)?
- What if there is constraint?
Preliminaries: optimization: Projected Gradient descent

- If objective function is
  - \( L\)-l.c.g., but not strongly convex
  - Constrained to convex set \( Q \)

- Projected gradient descent
  \[
  x_{k+1} = \Pi_Q \left( x_k - \frac{1}{L} \nabla f(x_k) \right) = \arg\min_{\hat{x} \in Q} \left\| \hat{x} - \left( x_k - \frac{1}{L} \nabla f(x_k) \right) \right\|
  = \arg\min_{x \in Q} f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \| x - x_k \|^2
  \]

- Rate of convergence: \( O \left( \frac{L}{\epsilon} \right) \)
  - Compare with Newton \( O \left( \sqrt{\frac{L}{\epsilon}} \right) \), interior point \( O \left( \log \frac{1}{\epsilon} \right) \)
Preliminaries: optimization: Projected Gradient descent

- Projected gradient descent

\[
x_{k+1} = \Pi_Q \left( x_k - \frac{1}{L} \nabla f(x_k) \right) = \arg\min_{\hat{x} \in Q} \left\| \hat{x} - \left( x_k - \frac{1}{L} \nabla f(x_k) \right) \right\|
\]

\[
= \arg\min_{x \in Q} f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2} \| x - x_k \|^2
\]

- Property 1: monotonic decreasing

\[
f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \| x_{k+1} - x_k \|^2 \quad \text{L-l.c.g.}
\]

\[
\leq f(x_k) + \langle \nabla f(x_k), x_k - x_k \rangle + \frac{L}{2} \| x_k - x_k \|^2 \quad \text{Def } x_{k+1}
\]

\[
= f(x_k)
\]
Preliminaries: optimization: Projected Gradient descent

- **Property 2:**

\[
\forall \mathbf{x} \in Q \quad \left\langle \mathbf{x} - \mathbf{x}_{k+1}, (\mathbf{x}_k - \frac{1}{L} \nabla f(\mathbf{x}_k)) - \mathbf{x}_{k+1} \right\rangle \leq 0
\]

\[
f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2} \left\| \mathbf{x}_{k+1} - \mathbf{x}_k \right\|^2
\]

**Property 2**

\[
\leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{L}{2} \left\| \mathbf{x} - \mathbf{x}_k \right\|^2 - \frac{L}{2} \left\| \mathbf{x} - \mathbf{x}_{k+1} \right\|^2 \quad \forall \mathbf{x} \in Q
\]

**Convexity of** \( f \)

\[
\leq f(\mathbf{x}) + \frac{L}{2} \left\| \mathbf{x} - \mathbf{x}_k \right\|^2 - \frac{L}{2} \left\| \mathbf{x} - \mathbf{x}_{k+1} \right\|^2 \quad \forall \mathbf{x} \in Q
\]
Preliminaries: optimization: Projected Gradient descent

- Put together

\[ f(x_{k+1}) \leq f(x) + \frac{L}{2} \|x - x_k\|^2 - \frac{L}{2} \|x - x_{k+1}\|^2 \quad \forall x \in Q \]

Let \( x = x^* \):

\[
0 \leq \frac{L}{2} \|x^* - x_{k+1}\|^2 \leq -\epsilon_{k+1} + \frac{L}{2} \|x^* - x_k\|^2
\]

\[
\leq \sum_{i=1}^{k+1} \epsilon_i + \frac{L}{2} \|x^* - x_0\|^2
\]

\[
\leq -(k+1)\epsilon_{k+1} + \frac{L}{2} \|x^* - x_0\|^2 \quad (\epsilon_k \text{ monotonic decreasing})
\]

\[
\epsilon_{k+1} \leq \frac{L}{2(k+1)} \|x^* - x_0\|^2
\]
Preliminaries: optimization:

Subgradient method

- Objective is continuous but not differentiable
- Subgradient method for \( \min_{x \in Q} f(x) \)

\[
x_{k+1} = \Pi_Q (x_k - s_k \nabla f(x_k))
\]

where \( \nabla f(x_k) \in \partial f(x_k) \) (arbitrary subgradient)

- Rate of convergence \( O \left( \frac{1}{\epsilon^2} \right) \)
- Summary

\[
O \left( \frac{1}{\epsilon^2} \right) \quad \Rightarrow \quad O \left( \frac{L}{\epsilon} \right) \quad \Rightarrow \quad \frac{\ln \frac{1}{\epsilon}}{-\ln(1 - \frac{\sigma}{L})}
\]

non-smooth \quad \quad \quad \quad L-l.c.g. \quad \quad \quad \quad L-l.c.g. \ & \sigma\text{-strongly convex}
Outline

- The problem from machine learning perspective
- Preliminaries
  - Convex analysis and gradient descent
- Nesterov’s optimal gradient method
  - Lower bound of optimization
  - Optimal gradient method
- Utilizing structure: composite optimization
  - Smooth minimization
  - Excessive gap minimization
- Conclusion
Optimal gradient method

Lower bound

- Consider the set of $L$-l.c.g. functions
  - For any $\epsilon > 0$, there exists an $L$-l.c.g. function $f$, such that any first-order method takes at least
    \[ k = O \left( \sqrt{\frac{L}{\epsilon}} \right) \]
    steps to ensure $\epsilon_k < \epsilon$.
  - First-order method means
    \[ x_k \in x_0 + \text{span} \{ \nabla f(x_0), \ldots, \nabla f(x_{k-1}) \} \]
  - Not saying: there exists an $L$-l.c.g. function $f$, such that for all $\epsilon > 0$ any first-order method takes at least $k = O(\sqrt{L/\epsilon})$ steps to ensure $\epsilon_k < \epsilon$.
  - Gap: recall the upper bound $O \left( \frac{L}{\epsilon} \right)$ of GD, two possibilities.
Optimal gradient method: Primitive Nesterov

- Problem under consideration

\[
\min_{w} f(w) \quad w \in Q
\]

where \( f \) is \( L \)-l.c.g., \( Q \) is convex

- Big results
  - He proposed an algorithm attaining \( \sqrt{L/\varepsilon} \)
  - **Not for free**: require an oracle to project a point onto \( Q \) in \( L_2 \) sense
Construct quadratic functions $\phi_k(x)$ and $\lambda_k > 0$

1. $\phi_k(x) = \phi_k^* + \frac{\gamma_k}{2} \| x - v_k \|^2$
2. $\exists x_k, s.t. f(x_k) \leq \phi_k^*$
3. $\phi_k(x) \leq (1 - \lambda_k)f(x) + \lambda_k \phi_0(x)$
4. $\lambda_k \to 0$

$f(x_k) \overset{2}{\leq} \phi_k^* \overset{1}{\leq} \phi_k(x^*)$

$\overset{3}{\leq} (1 - \lambda_k)f(x^*) + \lambda_k \phi_0(x^*)$

$f(x_k) - f(x^*) \leq \lambda_k(\phi_0(x^*) - f(x^*))$

$\to 0$
Construct quadratic functions $\phi_k(x)$ and $\lambda_k > 0$

1. $\phi_k(x) = \phi_k^* + \frac{\gamma_k}{2} \| x - v_k \|^2$
2. $\exists x_k, s.t. f(x_k) \leq \phi_k^*$
3. $\phi_k(x) \leq (1 - \lambda_k)f(x) + \lambda_k\phi_0(x)$
4. $\lambda_k \rightarrow 0$

$f(x_k) \overset{2}{\leq} \phi_k^* \overset{1}{\leq} \phi_k(x^*)$

$\overset{3}{\leq} (1 - \lambda_k)f(x^*) + \lambda_k\phi_0(x^*)$

$f(x_k) - f(x^*) \leq \lambda_k(\phi_0(x^*) - f(x^*)) \rightarrow 0$
Primitive Nesterov: Rate of convergence

Nesterov constructed, in a highly non-trivial way, the $\phi_k(x)$ and $\lambda_k$, s.t.

- $x_k$ has closed form (grad desc)
- $\lambda_k \leq \frac{4L}{(2\sqrt{L}+k\sqrt{\gamma_0})^2}$

Furthermore, if $f$ is $\sigma$-strongly convex, then

$$\lambda_k \leq \left(1 - \sqrt{\frac{\sigma}{L}}\right)^k$$

Rate of convergence sheerly depends on $\lambda_k$
Primitive Nesterov: Dealing with constraints

- $x_k$ has closed form by gradient descent
  \[ x_{k+1} = x_k - \gamma \nabla f(x_k) \]

- When constrained to set $Q$, modify by
  \[ x_{k+1}^Q = \Pi_Q (x_k - \gamma \nabla f(x_k)) = \arg\min_{x \in Q} \| x - (x_k - \gamma \nabla f(x_k)) \| \]

- New gradient:
  \[ g_k^Q := \gamma^{-1} \left( x_k - x_{k+1}^Q \right) \]

- This new gradient keeps all important properties of gradient, also keeping the rate of convergence
Primitive Nesterov: Gradient mapping

- $x_k$ has closed form by gradient descent
  
  $$x_{k+1} = x_k - \gamma \nabla f(x_k)$$

- When constrained to set $Q$, modify by
  
  $$x_{k+1}^Q = \Pi_Q (x_k - \gamma \nabla f(x_k)) = \arg\min_{x \in Q} \|x - (x_k - \gamma \nabla f(x_k))\|$$

- New gradient:
  
  $$g_k^Q := \gamma^{-1} \left(x_k - x_{k+1}^Q\right)$$

- This new gradient keeps all important properties of gradient, also keeping the rate of convergence
Primitive Nesterov

- **Summary**

\[ \min_w f(w) \quad \text{w} \in Q \]

where \( f \) is \( L\)-l.c.g., \( Q \) is convex.

- **Rate of convergence**

\[ \sqrt{\frac{L}{\epsilon}} \quad \text{if no strong convexity} \]

\[ \frac{\ln \frac{1}{\epsilon}}{\ln \left(1 - \frac{\sigma}{L}\right)} \quad \text{if } \sigma\text{-strongly convexity} \]
Primitive Nesterov: Example

$$\min_{x,y} \frac{1}{2} x^2 + 2y^2$$
$$\mu = 1, \quad L = 4$$

$$\min_{x \geq 0, y \geq 0} \frac{1}{2} (x + y)^2$$
$$\mu = 0, \quad L = 2$$
Extension: Non-Euclidean norm

- Remember strong convexity and l.c.g. are wrt some norm
  - We have implicitly used Euclidean norm ($L_2$ norm)
  - Some functions are strongly convex wrt other norms
  - Negative entropy $\sum_i x_i \ln x_i$ is
    - Not l.c.g. wrt $L_2$ norm
    - l.c.g. wrt $L_1$ norm $\|x\|_1 = \sum_i x_i$
    - strongly convex wrt $L_1$ norm.

Can Nesterov’s approach be extended to non-Euclidean norm?
Extension: Non-Euclidean norm

- Remember strong convexity and \( l.c.g. \) are wrt some norm
  - We have implicitly used Euclidean norm (\( L_2 \) norm)
  - Some functions are \( l.c.g. \) wrt other norms
  - Negative entropy \( \sum_i x_i \ln x_i \) is
    - Not \( l.c.g. \) wrt \( L_2 \) norm
    - \( l.c.g. \) wrt \( L_1 \) norm \( \|x\|_1 = \sum_i x_i \)
    - strongly convex wrt \( L_1 \) norm.

Can Nesterov’s approach be extended to non-Euclidean norm?

Yes

\[ \text{\( k \times k \)} \]
Extension: Non-Euclidean norm

Suppose the objective function \( f \) is l.c.g. wrt \( \| \cdot \| \).

Use a prox-function \( d \) on \( Q \) which is \( \sigma \)-strongly convex wrt \( \| \cdot \| \), and

\[
\min_{x \in Q} d(x) = 0 \quad D := \max_{x \in Q} d(x)
\]

### Algorithm 1: Nesterov's algorithm for non-Euclidean norm

**Output:** A sequence \( \{y^k\} \) converging to the optimal at \( O(1/k^2) \) rate.

1. **Initialize:** Set \( x^0 \) to a random value in \( Q \).
2. **for** \( k = 0, 1, 2, \ldots \) **do**
3. Query the gradient of \( f \) at point \( x^k \): \( \nabla f(x^k) \).
4. Find \( y^k \leftarrow \arg\min_{x \in Q} \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{2L} \|x - x^k\|^2 \).
5. Find \( z^k \leftarrow \arg\min_{x \in Q} \frac{L}{\sigma} d(x) + \sum_{i=0}^{k} \frac{i+1}{2} \left\langle \nabla f(x^i), x - x^i \right\rangle .
6. Update \( x^{k+1} \leftarrow \frac{2}{k+3} z^k + \frac{k+1}{k+3} y^k \).
**Extension:**

**Non-Euclidean norm**

Suppose the objective function $f$ is l.c.g. wrt $\| \cdot \|$. Use a prox-function $d$ on $Q$ which is $\sigma$-strongly convex wrt $\| \cdot \|$, and

\[
\min_{x \in Q} d(x) = 0 \quad D := \max_{x \in Q} d(x)
\]

**Algorithm 1:** Nesterov's algorithm for non-Euclidean norm

**Output:** A sequence $\{ y^k \}$ converging to the optimal at $O(1/k^2)$ rate.

1. Initialize: Set $x^0$ to a random value in $Q$.
2. for $k = 0, 1, 2, \ldots$ do
3. Query the gradient of $f$ at point $x^k$: $\nabla f(x^k)$.
4. Find $y^k \leftarrow \arg\min_{x \in Q} \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2}L \| x - x^k \|^2$.
5. Find $z^k \leftarrow \arg\min_{x \in Q} \frac{L}{\sigma}d(x) + \sum_{i=0}^{k} \frac{i+1}{2} \langle \nabla f(x^i), x - x^i \rangle$.
6. Update $x^{k+1} \leftarrow \frac{2}{k+3} z^k + \frac{k+1}{k+3} y^k$. 

I won’t mention details.
Extension: Non-Euclidean norm

- Rate of convergence

\[ f(y_k) - f(x^*) \leq \frac{4Ld(x^*)}{\sigma(k + 1)(k + 2)} \]

- Applications will be given later.
Immediate application: Non-smooth functions

- Objective function not differentiable
  - Suppose it is the Fenchel dual of some function $f$
    \[
    \min_x f^*(x) \quad \text{where } f \text{ is defined on } Q
    \]

- Idea: smooth the non-smooth function.
  - Add a small $\sigma$-strongly convex function $d$ to $f$

\[
\begin{align*}
  f + d \text{ is } \sigma\text{-strongly convex} & \quad \rightarrow \quad (f + d)^* \text{ is } \frac{1}{\sigma}\text{-l.c.g}
\end{align*}
\]
Immediate application:
Non-smooth functions

- $(f + \epsilon d)^*(x)$ approximates $f^*(x)$
- If $0 \leq d(u) \leq D$ for $u \in Q$ then

$$f^*(x) - \epsilon D \leq (f + \epsilon d)^*(x) \leq f^*(x)$$

Proof

$$\max_u \langle u, x \rangle - f(u) - \epsilon D \leq \max_u \langle u, x \rangle - f(u) - \epsilon d(u) \leq \max_u \langle u, x \rangle - f(u) - 0$$

$$f^*(x) - \epsilon D \leq (f + \epsilon d)^*(x) \leq f^*(x)$$
Immediate application: Non-smooth functions

- \((f + \epsilon d)^*(x)\) approximates \(f^*(x)\) well
  - If \(d(u) \in [0, D]\) on \(Q\), then \((f + \epsilon d)^*(x) - f^*(x) \in [-\epsilon D, 0]\)

- Algorithm (given precision \(\epsilon\))
  - Fix \(\hat{\epsilon} = \frac{\epsilon}{2D}\)
  - Optimize \((f + \hat{\epsilon} d)^*(x)\) (l.c.g. function) to precision \(\epsilon/2\)

- Rate of convergence

\[
\sqrt{\frac{1}{\epsilon} L} = \sqrt{\frac{1}{\epsilon} \cdot \frac{1}{\hat{\epsilon} \sigma}} = \sqrt{\frac{2D}{\sigma \epsilon^2}} = \frac{1}{\epsilon} \sqrt{\frac{2D}{\sigma}}
\]
Outline

- The problem from machine learning perspective
- Preliminaries
  - Convex analysis and gradient descent
- Nesterov’s optimal gradient method
  - Lower bound of optimization
  - Optimal gradient method
- Utilizing structure: composite optimization
  - Smooth minimization
  - Excessive gap minimization
- Conclusion
Composite optimization

- Many applications have objectives in the form of

\[ J(w) = f(w) + g^*(Aw) \]

where

- \( f \) is convex on the region \( E_1 \) with norm \( \| \cdot \|_1 \)
- \( g \) is convex on the region \( E_2 \) with norm \( \| \cdot \|_2 \)

- Very useful in machine learning
  - \( Aw \) corresponds to linear model
Composite optimization

- Example: binary SVM

\[
J(w) = \frac{\lambda}{2} \|w\|^2 + \min_{b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} [1 - y_i(\langle x_i, w \rangle + b)]_+
\]

- \[ A = -(y_1 x_1, \ldots, y_n x_n)^\top \]
- \( g^* \) is the dual of \( g(\alpha) = -\sum_i \alpha_i \) over \( Q_2 = \{ \alpha \in [0, n^{-1}]^n : \sum_i y_i \alpha_i = 0 \} \)
Composite optimization 1: Smooth minimization

\[ J(w) = f(w) + g^*(Aw) \]

- Let us only assume that \( \nabla f \) is \( M \)-l.c.g wrt \( \| \cdot \|_1 \)

- Smooth \( g^* \) into \((g + \mu d_2)^*\) (\( d_2 \) is \( \sigma_2 \)-strongly convex wrt \( \| \cdot \|_2 \))

then \( J_\mu(w) = f(w) + (g + \mu d_2)^*(Aw) \)

is \( \left( M + \frac{1}{\mu \sigma_2} \| A \|_{1,2}^2 \right) \)-l.c.g

Apply Nesterov on \( J_\mu(w) \)
Composite optimization 1: Smooth minimization

- Rate of convergence
  - to find an $\epsilon$ accurate solution, it costs
    \[
    4 \| A \|_{1,2} \sqrt{\frac{D_1 D_2}{\sigma_1 \sigma_2}} \cdot \frac{1}{\epsilon} + \sqrt{MD_1 \over \sigma_1 \epsilon}
    \]
    steps.

  \[\begin{align*}
  d_1 & \text{ is } \sigma_1\text{-strongly convex wrt } \| \cdot \|_1 \\
  d_2 & \text{ is } \sigma_2\text{-strongly convex wrt } \| \cdot \|_2
  \end{align*}\]

  \[\begin{align*}
  D_1 & := \max_{w \in E_1} d_1(w) \quad D_2 := \max_{\alpha \in E_2} d_2(\alpha)
  \end{align*}\]
Composite optimization 1: Smooth minimization

- Example: matrix game

\[
\arg\min_{\mathbf{w} \in \Delta_n} \left\{ \langle \mathbf{c}, \mathbf{w} \rangle + \max_{\mathbf{\alpha} \in \Delta_m} \{ \langle A\mathbf{w}, \mathbf{\alpha} \rangle + \langle \mathbf{b}, \mathbf{\alpha} \rangle \} \right\}
\]

- Use Euclidean distance

\[
E_1 = \Delta_n \quad \| \mathbf{w} \|_1 = \left( \sum_i w_i^2 \right)^{1/2} \quad d_1(\mathbf{w}) = \frac{1}{2} \sum_i (w_i - n^{-1})^2 \quad \sigma_1 = \sigma_2 = 1
\]

\[
E_2 = \Delta_m \quad \| \mathbf{\alpha} \|_2 = \left( \sum_i \alpha_i^2 \right)^{1/2} \quad d_2(\mathbf{\alpha}) = \frac{1}{2} \sum_i (\alpha_i - m^{-1})^2 \quad D_1 < 1, \ D_2 < 1
\]

\[
\| A \|_{1,2}^2 = \lambda_{\max}^{1/2}(A^\top A)
\]

May scale with \( O(nm) \)

\[
f(\mathbf{w}_k) - f(\mathbf{w}^*) \leq \frac{4\lambda_{\max}(A^\top A)}{k + 1}
\]
Composite optimization 1: Smooth minimization

- Example: matrix game

\[
\begin{align*}
\arg\min_{\mathbf{w} \in \Delta_n} & \quad \langle \mathbf{c}, \mathbf{w} \rangle + \max_{\mathbf{\alpha} \in \Delta_m} \left\{ \langle \mathbf{A}\mathbf{w}, \mathbf{\alpha} \rangle + \langle \mathbf{b}, \mathbf{\alpha} \rangle \right\} \\
& \quad \text{subject to } f(\mathbf{w}) \text{ and } g^*(\mathbf{A}\mathbf{w})
\end{align*}
\]

- Use Entropy distance

\[
\begin{align*}
E_1 &= \Delta_n \quad \|\mathbf{w}\|_1 = \sum_i |w_i| \quad d_1(\mathbf{w}) = \ln n + \sum_i w_i \ln w_i \quad \sigma_1 = \sigma_2 = 1 \\
E_2 &= \Delta_m \quad \|\mathbf{\alpha}\|_2 = \sum_i |\alpha_i| \quad d_2(\mathbf{\alpha}) = \ln m + \sum_i \alpha_i \ln \alpha_i \\
\|\mathbf{A}\|_{1,2} &= \max_{i,j} |A_{i,j}| \\
f(\mathbf{w}_k) - f(\mathbf{w}^*) &\leq \frac{4 \left( \ln n \ln m \right)^{\frac{1}{2}}}{k + 1} \max_{i,j} \|A_{i,j}\| \\
D_1 &= \ln n \\
D_2 &= \ln m
\end{align*}
\]
Composite optimization 1: Smooth minimization

- Disadvantages:
  - Fix the smoothing beforehand using prescribed accuracy $\epsilon$
  - No convergence criteria because real min is unknown.
Composite optimization 2: Excessive gap minimization

- Primal-dual
  - Easily upper bounds the duality gap
- Idea
  - Assume objective function takes the form
    \[ J(w) = f(w) + g^*(Aw) \]
  - Utilizes the \textit{adjoint} form
    \[ D(\alpha) = -g(\alpha) - f^*(-A^\top \alpha) \]
- Relations:
  \[ \forall w, \alpha \quad J(w) \geq D(\alpha) \quad \text{and} \quad \inf_{w \in E_1} J(w) = \sup_{\alpha \in E_2} D(\alpha) \]
Composite optimization 2: Excessive gap minimization

- Sketch of idea
  - Assume $f$ is $L_f$-l.c.g. and $g$ is $L_g$-l.c.g.
  - Smooth both $f^*$ and $g^*$ by prox-functions $d_1, d_2$

$$J_{\mu_2}(w) = f(w) + (g + \mu_2 d_2)^*(A w)$$

$$D_{\mu_1}(\alpha) = -g(\alpha) - (f + \mu_1 d_1)^*(-A^T \alpha)$$
Composite optimization 2: Excessive gap minimization

Sketch of idea

Maintain two point sequences \( \{w_k\} \) and \( \{\alpha_k\} \) and two regularization sequences \( \{\mu_1(k)\} \) and \( \{\mu_2(k)\} \)

such that

\[
J_{\mu_2(k)}(w_k) \leq D_{\mu_1(k)}(\alpha_k)
\]

\[
\mu_1(k) \to 0 \quad \mu_2(k) \to 0
\]
Composite optimization 2: Excessive gap minimization

\[
J_{\mu_2(k)}(w_k) \leq D_{\mu_1(k)}(\alpha_k)
\]

- **Challenge:**
  - How to efficiently find the initial point \(w_1, \alpha_1, \mu_1(1), \mu_2(1)\) that satisfy excessive gap condition.
  - Given \(w_k, \alpha_k, \mu_1(k), \mu_2(k)\), with new \(\mu_1(k+1)\) and \(\mu_2(k+1)\) how to efficiently find \(w_{k+1}\) and \(\alpha_{k+1}\).
  - How to anneal \(\mu_1(k)\) and \(\mu_2(k)\) (otherwise one step done).

- **Solution**
  - Gradient mapping
  - Bregman projection (very cool)
Composite optimization 2: Excessive gap minimization

- Rate of convergence:

\[
J(w_k) - D(\alpha_k) \leq \frac{4 \|A\|_{1,2}}{k + 1} \sqrt{\frac{D_1 D_2}{\sigma_1 \sigma_2}}
\]

- \(f\) is \(\sigma\)-strongly convex
  - No need to add prox-function to \(f\), \(\mu_1(k) \equiv 0\)

\[
J(w_k) - D(\alpha_k) \leq \frac{4 D_2}{\sigma_2 k(k + 1)} \left( \frac{\|A\|_{1,2}^2}{\sigma} + L_g \right)
\]
Composite optimization 2: Excessive gap minimization

- Example: binary SVM

\[
J(w) = \frac{\lambda}{2} \|w\|^2 + \min_{b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} [1 - y_i (\langle x_i, w \rangle + b)]_+ \]

- \( A = -(y_1 x_1, \ldots, y_n x_n)\)

- \( g^* \) is the dual of \( g(\alpha) = -\sum_i \alpha_i \) over
  \[
  E_2 = \{ \alpha \in [0, n^{-1}]^n : \sum_i y_i \alpha_i = 0 \}
  \]

- Adjoint form \( D(\alpha) = \sum_i \alpha_i - \frac{1}{2\lambda} \alpha^\top AA^\top \alpha \)
Composite optimization 2: Convergence rate for SVM

- Theorem: running on SVM for $k$ iterations

\[ J(w_k) - D(\alpha_k) \leq \frac{2L}{(k + 1)(k + 2)n} \]

- \[ L = \lambda^{-1} \|A\|^2 = \lambda^{-1} \|(y_1x_1, \ldots, y_nx_n)\|^2 \leq \frac{nR^2}{\lambda} \quad (\|x_i\| \leq R) \]

- Final conclusion

\[ J(w_k) - D(\alpha_k) \leq \varepsilon \quad \text{as long as} \quad k > O\left(\frac{R}{\sqrt{\lambda \varepsilon}}\right) \]
Composite optimization 2: Projection for SVM

- Efficient $O(n)$ time projection onto
  
  $$E_2 = \left\{ \alpha \in [0, n^{-1}]^n : \sum y_i \alpha_i = 0 \right\}$$

- Projection leads to a singly linear constrained QP
  
  $$\min_{\alpha} \sum_{i=1}^{n} (\alpha_i - m_i)^2$$
  
  $$s.t. \quad l_i \leq \alpha_i \leq u_i \quad \forall i \in [n];$$
  
  $$\sum_{i=1}^{n} \sigma_i \alpha_i = z.$$

Key tool:
Median finding takes $O(n)$ time
Automatic estimation of Lipschitz constant

- Automatic estimation of Lipschitz constant $L$
  - Geometric scaling
  - Does not affect the rate of convergence
Conclusion

- Nesterov’s method attains the lower bound
  - $O\left(\frac{L}{\epsilon}\right)$ for $L$-l.c.g. objectives
  - Linear rate for l.c.g. and strongly convex objectives

- Composite optimization
  - Attains the rate of the nice part of the function

- Handling constraints
  - Gradient mapping and Bregman projection
  - Essentially does not change the convergence rate

- Expecting wide applications in machine learning
  - Note: not in terms of generalization performance
Questions?