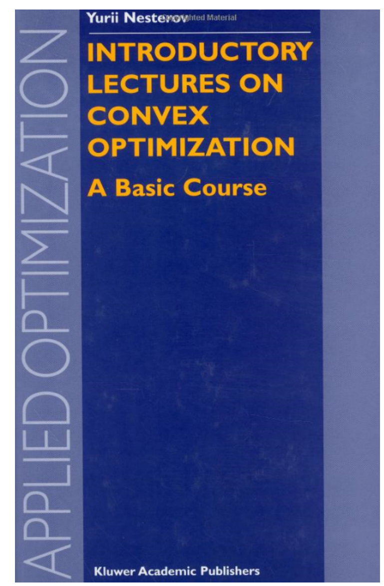
Yurii Nesterov
http://www.core.ucl.ac.be/~nesterov

# 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

# The problem

- Many machine learning problems have the form

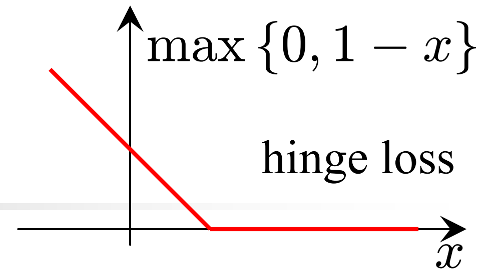
$$\min_{\mathbf{w}} J(\mathbf{w}) := \lambda \Omega(\mathbf{w}) + R_{\text{emp}}(\mathbf{w})$$

where

$$R_{\text{emp}}(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^{n} l(\mathbf{x}_i, \mathbf{y}_i; \mathbf{w})$$

  - $\mathbf{w}$ : weight vector
  - $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$ : training data
  - $l(\mathbf{x}, y; \mathbf{w})$ : convex and non-negative loss function
    - Can be non-smooth, possibly non-convex.
  - $\Omega(\mathbf{w})$ : convex and non-negative regularizer

# The problem: Examples




$$\min_{\mathbf{w}} \frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \xi_i$$

$$s.t. \quad \xi_i \geq 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \quad \forall 1 \leq i \leq n$$

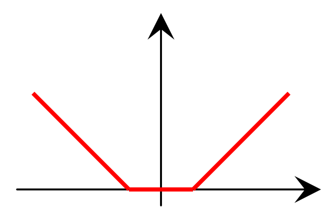
$$\xi_i \geq 0 \quad \forall 1 \leq i \leq n$$

$$\xi_i = \max\{0, 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle\}$$

$$\frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle\}$$

| Model (obj) | $\lambda\Omega(\mathbf{w})$ | + | $R_{\text{emp}}(\mathbf{w})$ |
|---|---|---|---|
| linear SVMs | $\frac{\lambda}{2}\|\mathbf{w}\|_2^2$ | + | $\frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i \langle \mathbf{w}, \mathbf{x}_i \rangle\}$ |
| $\ell_1$ logistic regression | $\lambda \|\mathbf{w}\|_1$ | + | $\frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle))$ |
| $\epsilon$-insensitive classify | $\frac{\lambda}{2}\|\mathbf{w}\|_2^2$ | + | $\frac{1}{n}\sum_{i=1}^{n} \max\{0, \|y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle\| - \epsilon\}$ |

$$\|\mathbf{w}_1\|_1 = \sum_i |w_i|$$



# The problem: More examples

Lasso
$$\underset{\mathbf{w}}{\operatorname{argmin}} \quad \lambda \cdot \|\mathbf{w}\|_1 + \|A\mathbf{w} - \mathbf{b}\|_2^2$$

Multi-task learning
$$\underset{\mathbf{w}}{\operatorname{argmin}} \quad \lambda \cdot \|W\|_{\mathrm{tr}} + \sum_{t=1}^{T} \|X_t \mathbf{w}_t - \mathbf{b}_t\|_2^2$$

$$\underset{\mathbf{w}}{\operatorname{argmin}} \quad \lambda \cdot \|W\|_{1,\infty} + \sum_{t=1}^{T} \|X_t \mathbf{w}_t - \mathbf{b}_t\|_2^2$$

Matrix game
$$\underset{\mathbf{w}\in\Delta_d}{\operatorname{argmin}} \quad \langle \mathbf{c}, \mathbf{w} \rangle + \max_{\mathbf{u}\in\Delta_n} \{\langle A\mathbf{w}, \mathbf{u}\rangle + \langle \mathbf{b}, \mathbf{u}\rangle\}$$

Entropy regularized LPBoost
$$\underset{\mathbf{w}\in\Delta_d}{\operatorname{argmin}} \quad \lambda\Delta(\mathbf{w}, \mathbf{w}^0) + \max_{\mathbf{u}\in\Delta_n} \langle A\mathbf{w}, \mathbf{u}\rangle$$

# The problem: Lagrange dual

Binary SVM

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

$$\text{s.t.} \qquad \alpha_i \in [0, n^{-1}]$$

$$\sum_i y_i \alpha_i = 0$$

where

$$Q_{ij} = y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

Entropy regularized LPBoost

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

$$\text{s.t.} \qquad \alpha_i \in [0, 1]$$

$$\sum_i \alpha_i = 1$$

# The problem

- Summary

$$\min_{\mathbf{w} \in Q} J(\mathbf{w})$$

where

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

- Solver: iterative method $\mathbf{w}_0, \ \mathbf{w}_1, \ \mathbf{w}_2, \ldots$
  - Want $\epsilon_k := J(\mathbf{w}_k) - J(\mathbf{w}^*)$ to decrease to 0 quickly

    where $\mathbf{w}^* := \operatorname{argmin}_{\mathbf{w} \in Q} J(\mathbf{w})$.

We only discuss optimization in this session, no generalization bound.

# The problem: What makes a good optimizer?

- Find an $\epsilon$ -approximate solution $\mathbf{w}_k$

$$J(\mathbf{w}_k) \leq \min_{\mathbf{w}} J(\mathbf{w}) + \epsilon$$

- 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:

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

- Use interchangeably:
  - Fix step index $k$, upper bound $\min_{1\le t\le k} \epsilon_t$
  - Fix precision $\epsilon$, how many steps needed for $\min_{1\le t\le k} \epsilon_t < \epsilon$
    - E.g. $\frac{1}{\epsilon^2}, \quad \frac{1}{\epsilon}, \quad \frac{1}{\sqrt{\epsilon}}, \quad \log\frac{1}{\epsilon}, \quad \log\log\frac{1}{\epsilon}$

# The problem: Collection of results

- Convergence rate:

| Objective function | Smooth | Smooth and very convex |
|---|---|---|
| Gradient descent | $O\left(\frac{1}{\epsilon}\right)$ | $O\left(\log\frac{1}{\epsilon}\right)$ |
| Nesterov | $O\left(\sqrt{\frac{1}{\epsilon}}\right)$ | $O\left(\log\frac{1}{\epsilon}\right)$ |
| Lower bound | $O\left(\sqrt{\frac{1}{\epsilon}}\right)$ | $O\left(\log\frac{1}{\epsilon}\right)$ |

- Composite non-smooth

Smooth + (dual of smooth)

$$O\left(\frac{1}{\epsilon}\right)$$

(very convex) + (dual of smooth)

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

# Outline

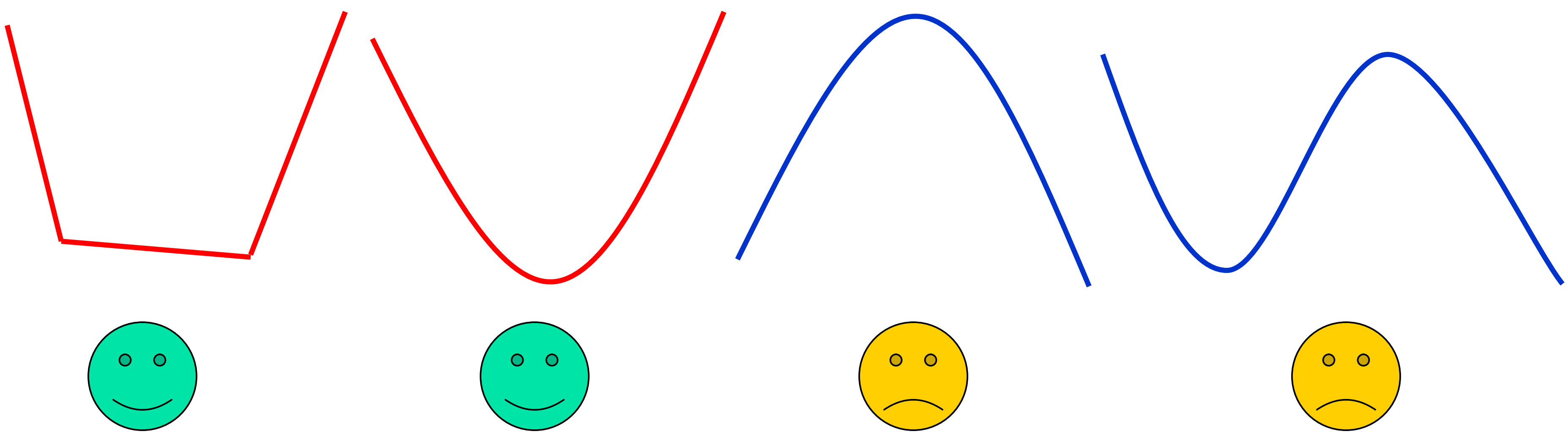
- The problem from machine learning perspective
- Preliminaries
  - Convex analysis and gradient descent
- Nesterov's optimal gradient method
  - Lower bound of optimization
  - Optimal gradient method
- Utilizing structure: composite optimization
  - Smooth minimization
  - Excessive gap minimization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

# Preliminaries: convex analysis Convex functions

- A function $f$ is convex iff

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

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

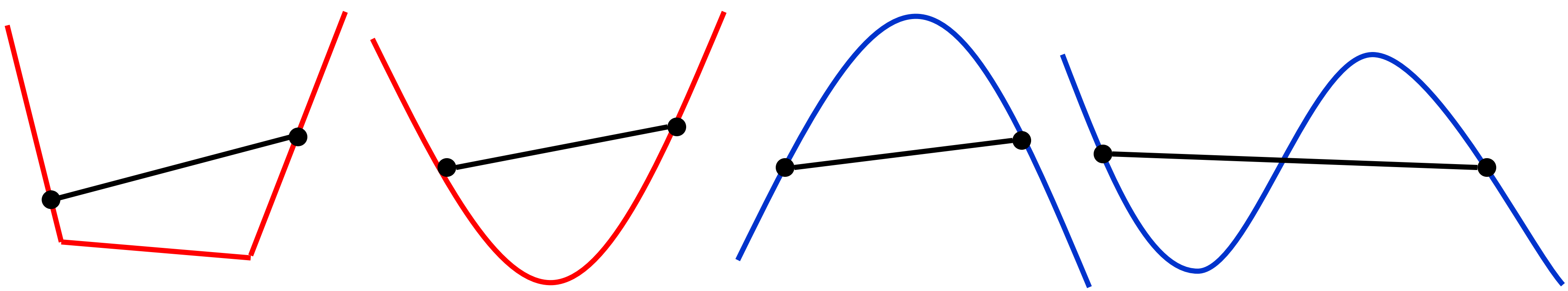


# Preliminaries: convex analysis Convex functions

- A function $f$ is convex iff

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

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y})$$



# Preliminaries: convex analysis
# Strong convexity

- A function $f$ is called $\sigma$-strongly convex wrt a norm $\|\cdot\|$ iff

$$f(\mathbf{x}) - \frac{1}{2}\sigma\|\mathbf{x}\|^2 \quad \text{is convex}$$

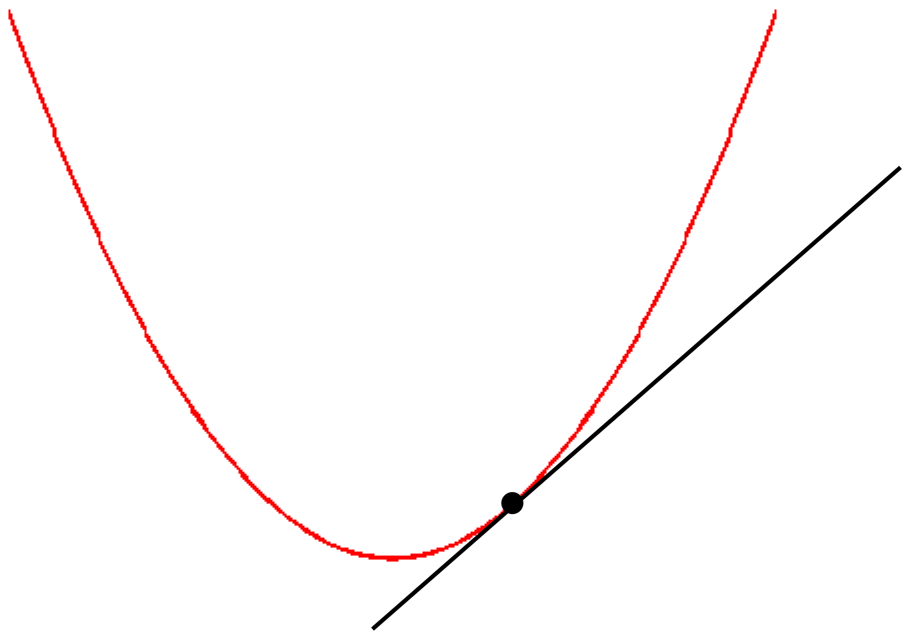
$\forall\ \mathbf{x}, \mathbf{y}, \lambda \in (0,1)$

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \le \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}) - \sigma \cdot \frac{\lambda(1-\lambda)}{2}\|\mathbf{x}-\mathbf{y}\|^2$$

# Preliminaries: convex analysis Strong convexity

- First order equivalent condition

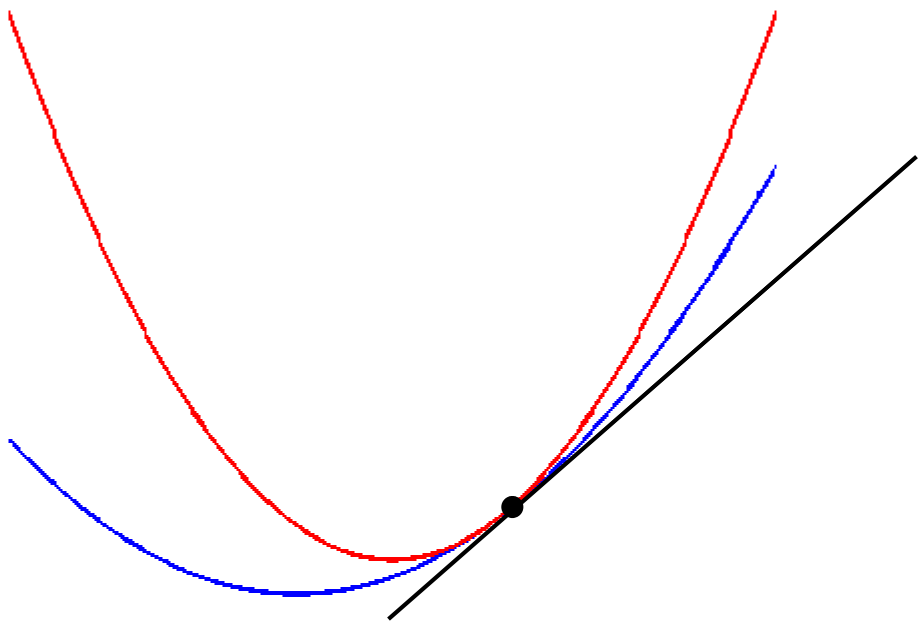
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\sigma}{2} \|\mathbf{x} - \mathbf{y}\|^2 \qquad \forall\ \mathbf{x}, \mathbf{y}$$



# Preliminaries: convex analysis Strong convexity

- First order equivalent condition

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\sigma}{2} \|\mathbf{x} - \mathbf{y}\|^2 \qquad \forall\ \mathbf{x}, \mathbf{y}$$



# Preliminaries: convex analysis Strong convexity

- Second order

$$\langle \nabla^2 f(\mathbf{x})\mathbf{y}, \mathbf{y} \rangle \geq \sigma \|\mathbf{y}\|^2 \qquad \forall\ \mathbf{x}, \mathbf{y}$$

  - If $\|\cdot\|$ Euclidean norm, then

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

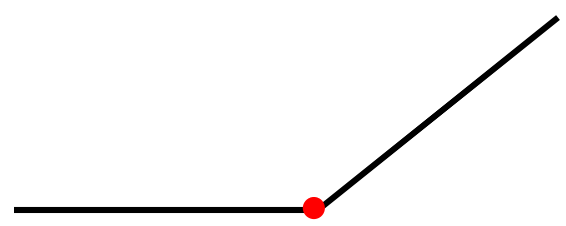
  - Lower bounds rate of change of gradient

# Preliminaries: convex analysis Lipschitz continuous gradient

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

$\exists L > 0$

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L \, \|\mathbf{x} - \mathbf{y}\| \qquad \forall \, \mathbf{x}, \mathbf{y}$$

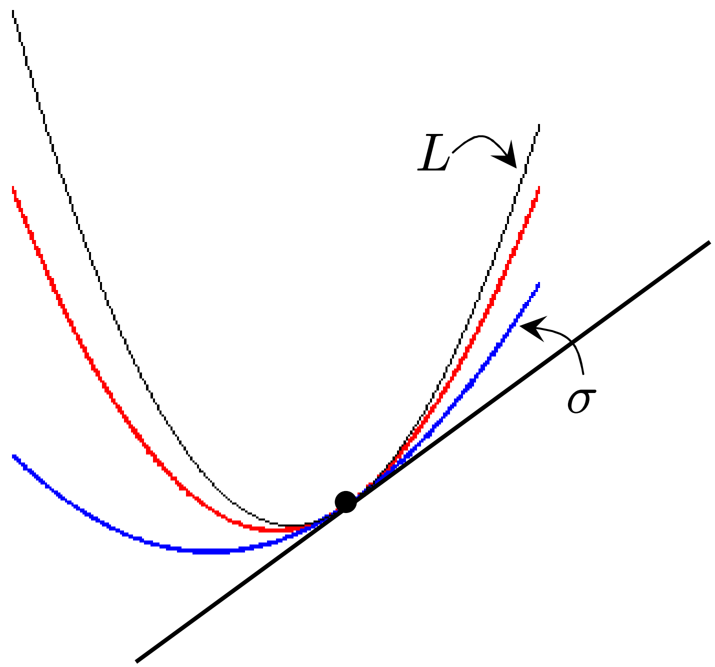


# Preliminaries: convex analysis Lipschitz continuous gradient

- Gradient is Lipschitz continuous (must be differentiable)

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \qquad \forall\ \mathbf{x}, \mathbf{y}$$

$$\Longleftrightarrow \quad f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 \qquad \forall\ \mathbf{x}, \mathbf{y}$$




$L$-$l.c.g$

# Preliminaries: convex analysis Lipschitz continuous gradient

- Gradient is Lipschitz continuous (must be differentiable)

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \qquad \forall\ \mathbf{x}, \mathbf{y}$$

$$\Longleftrightarrow \quad f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 \qquad \forall\ \mathbf{x}, \mathbf{y}$$

$$\Longleftrightarrow \quad \left\langle \nabla^2 f(\mathbf{x})\mathbf{y}, \mathbf{y} \right\rangle \leq L \|\mathbf{y}\|^2 \qquad \forall\ \mathbf{x}, \mathbf{y}$$

$$\nabla^2 f(x) \preceq L\mathbb{I} \qquad \text{if } L_2 \text{ norm}$$

# Preliminaries: convex analysis Fenchel Dual

- Fenchel dual of a function $f$

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

- Properties

$f^{\star\star} = f$ if $f$ is convex and closed

| $f$ | | $f^\star$ |
|---|---|---|
| $\sigma$ strongly convex | $\Leftrightarrow$ | $\frac{1}{\sigma}$-$l.c.g$ on $\mathbb{R}^d$ |
| $L$-$l.c.g$ on $\mathbb{R}^d$ | $\Leftrightarrow$ | $\frac{1}{L}$ strongly convex |

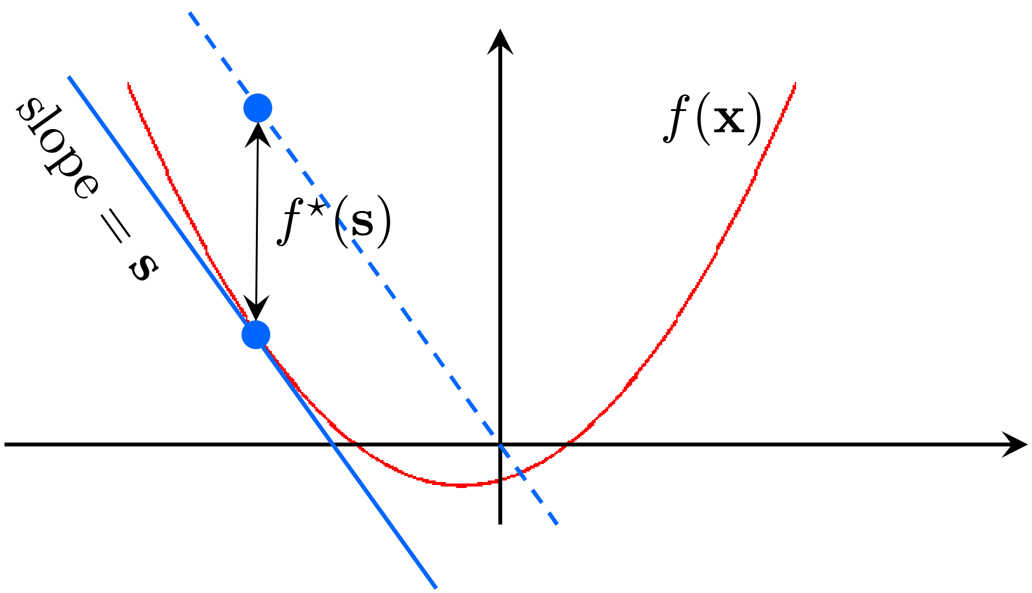
# Preliminaries: convex analysis Fenchel Dual

- Fenchel dual of a function $f$

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

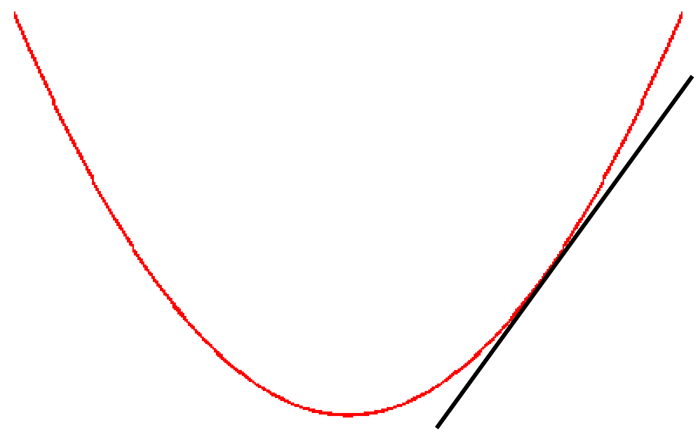
$$\mathbf{s} = \nabla f(\mathbf{x})$$

$$\mathbf{s} \in \partial f(\mathbf{x})$$



# Preliminaries: convex analysis: Subgradient

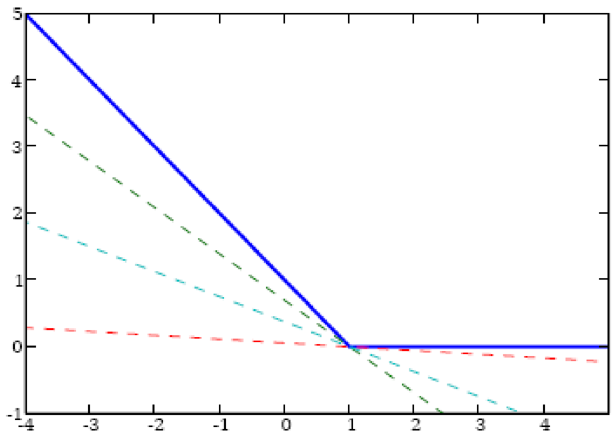
- Generalize gradient to non-differentiable functions
  - Idea: tangent plane lying below the graph of $f$



# Preliminaries: convex analysis: Subgradient

- Generalize gradient to non-differentiable functions
  - $\boldsymbol{\mu}$ is called a subgradient of $f$ at $\mathbf{x}$ if

$$f(\mathbf{x}') \geq f(\mathbf{x}) + \langle \mathbf{x}' - \mathbf{x}, \boldsymbol{\mu} \rangle \quad \forall \mathbf{x}'$$

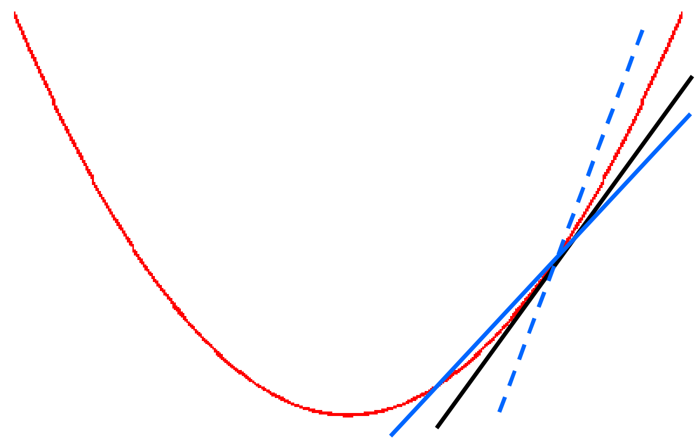



  - All such $\boldsymbol{\mu}$ comprise the subdifferential of $f$ at $\mathbf{x}$: $\partial f(\mathbf{x})$

# Preliminaries: convex analysis: Subgradient

- Generalize gradient to non-differentiable functions
  - $\boldsymbol{\mu}$ is called a subgradient of $f$ at $\mathbf{x}$ if

$$f(\mathbf{x}') \geq f(\mathbf{x}) + \langle \mathbf{x}' - \mathbf{x}, \boldsymbol{\mu} \rangle \quad \forall \mathbf{x}'$$



  - All such $\boldsymbol{\mu}$ comprise the subdifferential of $f$ at $\mathbf{x}$: $\partial f(\mathbf{x})$
  - Unique if $f$ is differentiable at $\mathbf{x}$

# Preliminaries: optimization: Gradient descent

- Gradient descent

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \nabla f(\mathbf{x}_k) \qquad s_k \geq 0$$

- Suppose $f$ is both $\sigma$-strongly convex and $L$-*l.c.g.*

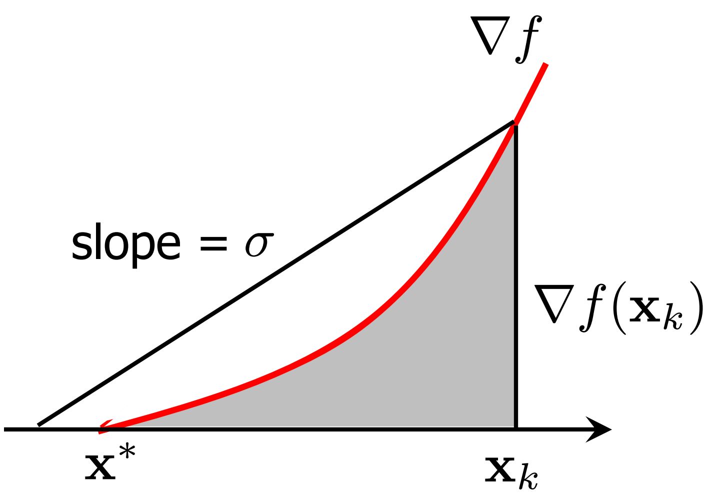
$$\epsilon_k := f(\mathbf{x}_k) - f(\mathbf{w}^*) \qquad \boxed{\epsilon_k \leq \left(1 - \frac{\sigma}{L}\right)^k \epsilon_0}$$

- 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

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)$$




shaded area $\leq$ triangle area

$$f(\mathbf{x}_k) - f(x^*) \qquad \frac{1}{2\sigma}\left\|\nabla f(\mathbf{x}_k)\right\|^2$$

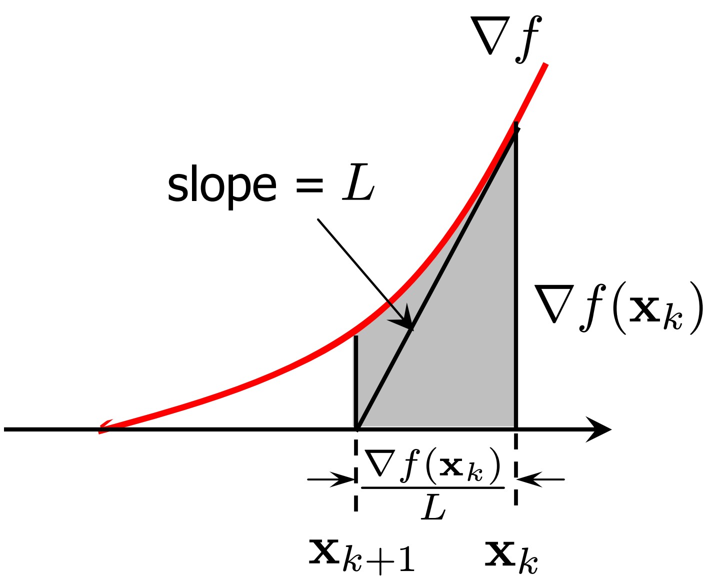
So

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{1}{2\sigma}\left\|\nabla f(\mathbf{x}_k)\right\|^2$$

# Preliminaries: optimization: Gradient descent

- Lower bound progress at each step

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)$$




shaded area $\geq$ triangle area

$\|$ $\|$

$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})$ $\quad$ $\frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2$

So

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|^2$$

# Preliminaries: optimization: Gradient descent

- Putting things together

distance to optimal progress

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

$$\Downarrow$$

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \left(1 - \frac{\sigma}{L}\right)(f(\mathbf{x}_k) - f(\mathbf{x}^*))$$

# Preliminaries: optimization: Gradient descent

- Putting things together

distance to optimal　　　　　　　　progress

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

⇓

$$\underbrace{f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)}_{\epsilon_{k+1}} \leq \left(1 - \frac{\sigma}{L}\right) \underbrace{(f(\mathbf{x}_k) - f(\mathbf{x}^*))}_{\epsilon_k}$$

# Preliminaries: optimization: Gradient descent

- Putting things together

distance to optimal — progress

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

⇓

$$\underbrace{f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)}_{\epsilon_{k+1}} \leq \left(1 - \frac{\sigma}{L}\right) \underbrace{(f(\mathbf{x}_k) - f(\mathbf{x}^*))}_{\epsilon_k}$$

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

$$\mathbf{x}_{k+1} = \Pi_Q\left(\mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)\right) = \operatorname*{argmin}_{\hat{\mathbf{x}}\in Q}\left\|\hat{\mathbf{x}} - (\mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k))\right\|$$

$$= \operatorname*{argmin}_{\mathbf{x}\in Q} f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k\rangle + \tfrac{L}{2}\left\|\mathbf{x} - \mathbf{x}_k\right\|^2$$

- 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

$$\mathbf{x}_{k+1} = \Pi_Q\left(\mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)\right) = \operatorname*{argmin}_{\hat{\mathbf{x}}\in Q}\left\|\hat{\mathbf{x}} - (\mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k))\right\|$$

$$= \operatorname*{argmin}_{\mathbf{x}\in Q} f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k\rangle + \tfrac{L}{2}\left\|\mathbf{x} - \mathbf{x}_k\right\|^2$$

- Property 1: monotonic decreasing

$$f(\mathbf{x}_{k+1}) \le f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle + \frac{L}{2}\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|^2 \quad \textit{L-l.c.g.}$$

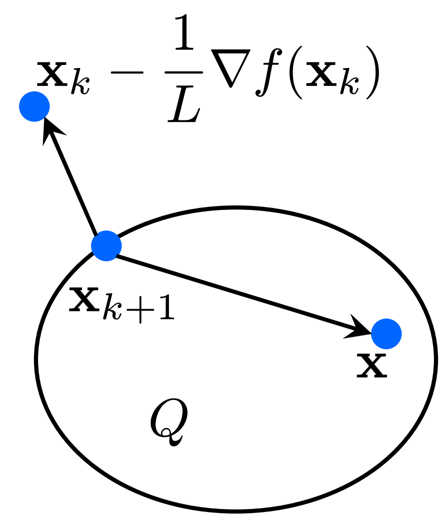
$$\le f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}_k\rangle + \frac{L}{2}\left\|\mathbf{x}_k - \mathbf{x}_k\right\|^2 \quad \text{Def } \mathbf{x}_{k+1} \text{ projection}$$

$$= f(\mathbf{x}_k)$$

# Preliminaries: optimization: Projected Gradient descent

- Property 2:

$$\forall\, \mathbf{x} \in Q \qquad \left\langle \mathbf{x} - \mathbf{x}_{k+1}, \left(\mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)\right) - \mathbf{x}_{k+1} \right\rangle \le 0$$




*L-l.c.g.*

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

Property 2

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

Convexity of $f$

$$\le f(\mathbf{x}) + \frac{L}{2}\left\|\mathbf{x} - \mathbf{x}_k\right\|^2 - \frac{L}{2}\left\|\mathbf{x} - \mathbf{x}_{k+1}\right\|^2 \qquad \forall\, \mathbf{x} \in Q$$

# Preliminaries: optimization: Projected Gradient descent

- Put together

$$f(\mathbf{x}_{k+1}) \le f(\mathbf{x}) + \frac{L}{2}\|\mathbf{x}-\mathbf{x}_k\|^2 - \frac{L}{2}\|\mathbf{x}-\mathbf{x}_{k+1}\|^2 \qquad \forall\, \mathbf{x} \in Q$$

Let $\mathbf{x} = \mathbf{x}^*$:

$$0 \le \frac{L}{2}\|\mathbf{x}^*-\mathbf{x}_{k+1}\|^2 \le -\epsilon_{k+1} + \frac{L}{2}\|\mathbf{x}^*-\mathbf{x}_k\|^2$$

$$\le \ldots \le \sum_{i=1}^{k+1}\epsilon_i + \frac{L}{2}\|\mathbf{x}^*-\mathbf{x}_0\|^2$$

$$\le -(k+1)\epsilon_{k+1} + \frac{L}{2}\|\mathbf{x}^*-\mathbf{x}_0\|^2 \qquad (\epsilon_k \text{ monotonic decreasing})$$

$$\Rightarrow \quad \epsilon_{k+1} \le \frac{L}{2(k+1)}\|\mathbf{x}^*-\mathbf{x}_0\|^2$$

# Preliminaries: optimization: Subgradient method

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

$$\mathbf{x}_{k+1} = \Pi_Q \left( \mathbf{x}_k - s_k \nabla f(\mathbf{x}_k) \right)$$

where $\nabla f(\mathbf{x}_k) \in \partial f(\mathbf{x}_k)$ (arbitrary subgradient)

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

| $O\left(\frac{1}{\epsilon^2}\right)$ | $O\left(\frac{L}{\epsilon}\right)$ | $\frac{\ln \frac{1}{\epsilon}}{-\ln(1 - \frac{\sigma}{L})}$ |
|---|---|---|
| non-smooth | *L-l.c.g.* | *L-l.c.g.* & $\sigma$-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

    $$\mathbf{x}_k \in \mathbf{x}_0 + \operatorname{span}\{\nabla f(\mathbf{x}_0), \ldots, \nabla f(\mathbf{x}_{k-1})\}$$

  - 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

$$\min_{\mathbf{w}} f(\mathbf{w}) \qquad \mathbf{w} \in Q$$

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

# Primitive Nesterov

Construct quadratic functions $\phi_k(\mathbf{x})$ and $\lambda_k > 0$

① $\phi_k(\mathbf{x}) = \phi_k^* + \frac{\gamma_k}{2}\|\mathbf{x} - \mathbf{v}_k\|^2$

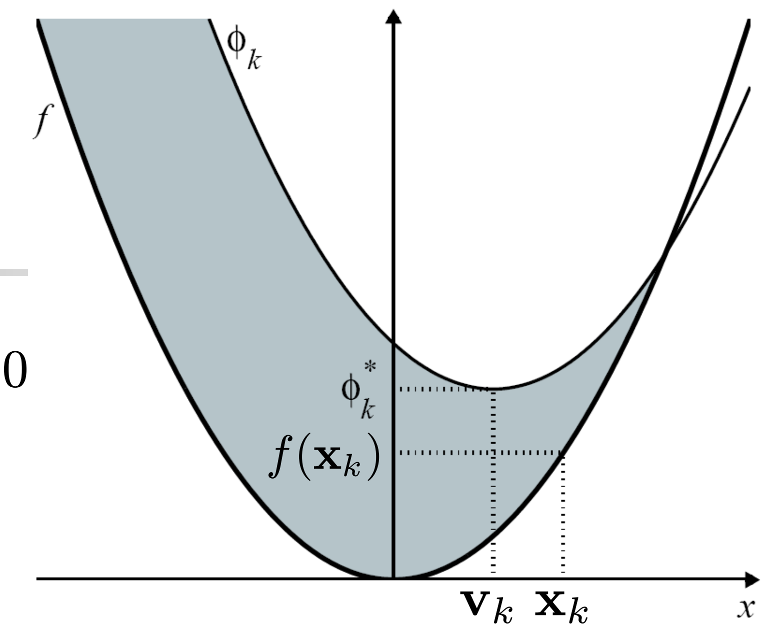
② $\exists\ \mathbf{x}_k, s.t.\ f(\mathbf{x}_k) \le \phi_k^*$

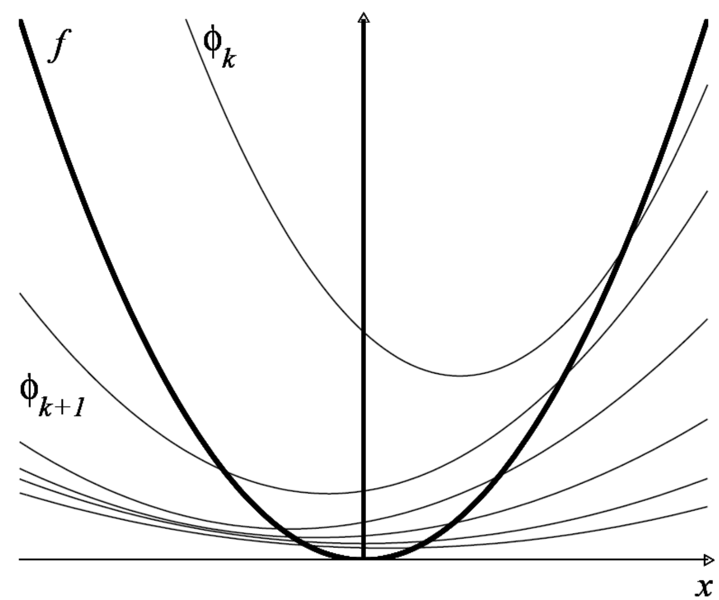
③ $\phi_k(\mathbf{x}) \le (1-\lambda_k) f(\mathbf{x}) + \lambda_k \phi_0(\mathbf{x})$

④ $\lambda_k \to 0$

$$\Rightarrow \quad f(\mathbf{x}_k) \overset{②}{\le} \phi_k^* \overset{①}{\le} \phi_k(\mathbf{x}^*) \overset{③}{\le} (1-\lambda_k) f(\mathbf{x}^*) + \lambda_k \phi_0(\mathbf{x}^*)$$

$$\Rightarrow \quad f(\mathbf{x}_k) - f(\mathbf{x}^*) \le \lambda_k(\phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)) \to 0$$



# Primitive Nesterov

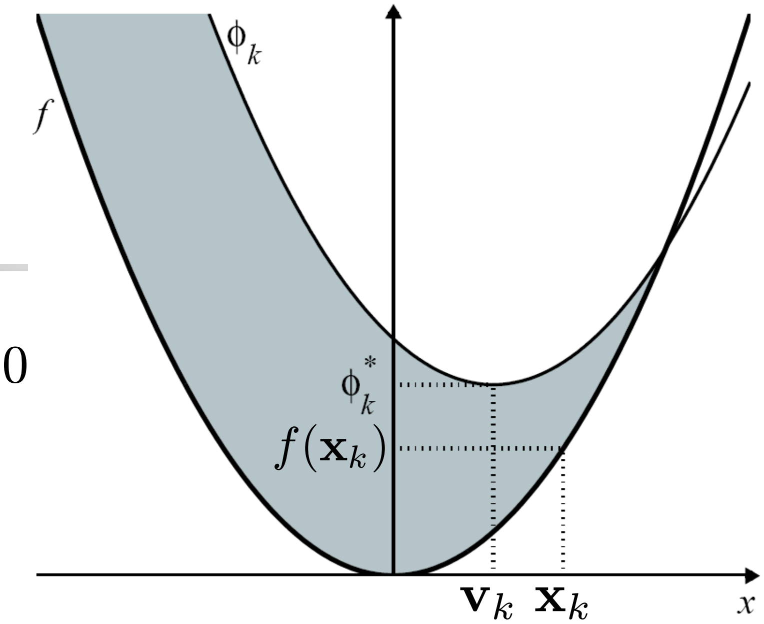
Construct quadratic functions $\phi_k(\mathbf{x})$ and $\lambda_k > 0$

① $$\phi_k(\mathbf{x}) = \phi_k^* + \frac{\gamma_k}{2}\|\mathbf{x} - \mathbf{v}_k\|^2$$

② $$\exists\ \mathbf{x}_k, s.t.\ f(\mathbf{x}_k) \le \phi_k^*$$

③ $$\phi_k(\mathbf{x}) \le (1-\lambda_k) f(\mathbf{x}) + \lambda_k \phi_0(\mathbf{x})$$

④ $$\lambda_k \to 0$$

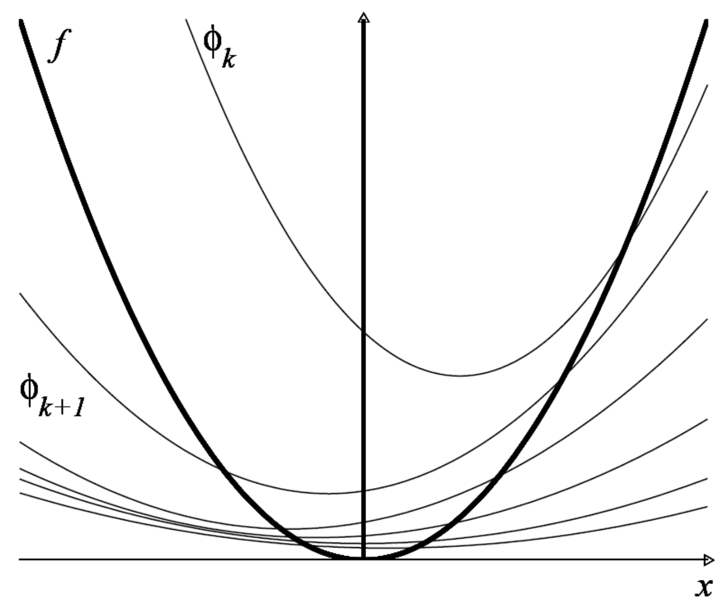



$$\Rightarrow\quad f(\mathbf{x}_k) \overset{②}{\le} \phi_k^* \overset{①}{\le} \phi_k(\mathbf{x}^*)$$

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

$$\Rightarrow\quad f(\mathbf{x}_k) - f(\mathbf{x}^*) \le \lambda_k(\phi_0(\mathbf{x}^*) - f(\mathbf{x}^*))$$

$$\to 0$$



# Primitive Nesterov: Rate of convergence

① $\phi_k(\mathbf{x}) = \phi_k^* + \frac{\gamma_k}{2}\|\mathbf{x} - \mathbf{v}_k\|^2$

② $\exists\ \mathbf{x}_k, s.t.\ f(\mathbf{x}_k) \leq \phi_k^*$

③ $\phi_k(\mathbf{x}) \leq (1-\lambda_k)f(\mathbf{x}) + \lambda_k \phi_0(\mathbf{x})$

④ $\lambda_k \to 0$

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \lambda_k(\phi_0(\mathbf{x}^*) - f(\mathbf{x}^*))$$

Rate of convergence sheerly depends on $\lambda_k$

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

- ✓ $\mathbf{x}_k$ 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

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

# Primitive Nesterov: Dealing with constraints

- $\mathbf{x}_k$ has closed form by gradient descent

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma \nabla f(\mathbf{x}_k)$$

- When constrained to set $Q$, modify by

$$\mathbf{x}_{k+1}^Q = \Pi_Q\left(\mathbf{x}_k - \gamma \nabla f(\mathbf{x}_k)\right) = \underset{\mathbf{x} \in Q}{\operatorname{argmin}} \left\|\mathbf{x} - (\mathbf{x}_k - \gamma \nabla f(\mathbf{x}_k))\right\|$$

  - New gradient:

$$\boldsymbol{g}_k^Q := \gamma^{-1}\left(\mathbf{x}_k - \mathbf{x}_{k+1}^Q\right)$$

gradient mapping

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

# Primitive Nesterov: Gradient mapping

- $\mathbf{x}_k$ has closed form by gradient descent

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma \nabla f(\mathbf{x}_k)$$

- When constrained to set $Q$ , modify by

$$\mathbf{x}_{k+1}^Q = \Pi_Q\left(\mathbf{x}_k - \gamma \nabla f(\mathbf{x}_k)\right) = \operatorname*{argmin}_{\mathbf{x} \in Q} \|\mathbf{x} - (\mathbf{x}_k - \gamma \nabla f(\mathbf{x}_k))\|$$

Expensive?

  - New gradient:

$$\boldsymbol{g}_k^Q := \gamma^{-1}\left(\mathbf{x}_k - \mathbf{x}_{k+1}^Q\right)$$

gradient mapping

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

# Primitive Nesterov

- Summary

$$\min_{\mathbf{w}} f(\mathbf{w}) \qquad \mathbf{w} \in Q$$

where $f$ is $L$-*l.c.g.*, $Q$ is convex.

- Rate of convergence

$$\sqrt{\frac{L}{\epsilon}}$$ if no strong convexity

$$\frac{\ln \frac{1}{\epsilon}}{-\ln(1-\frac{\sigma}{L})}$$ if $\sigma$-strongly convexity

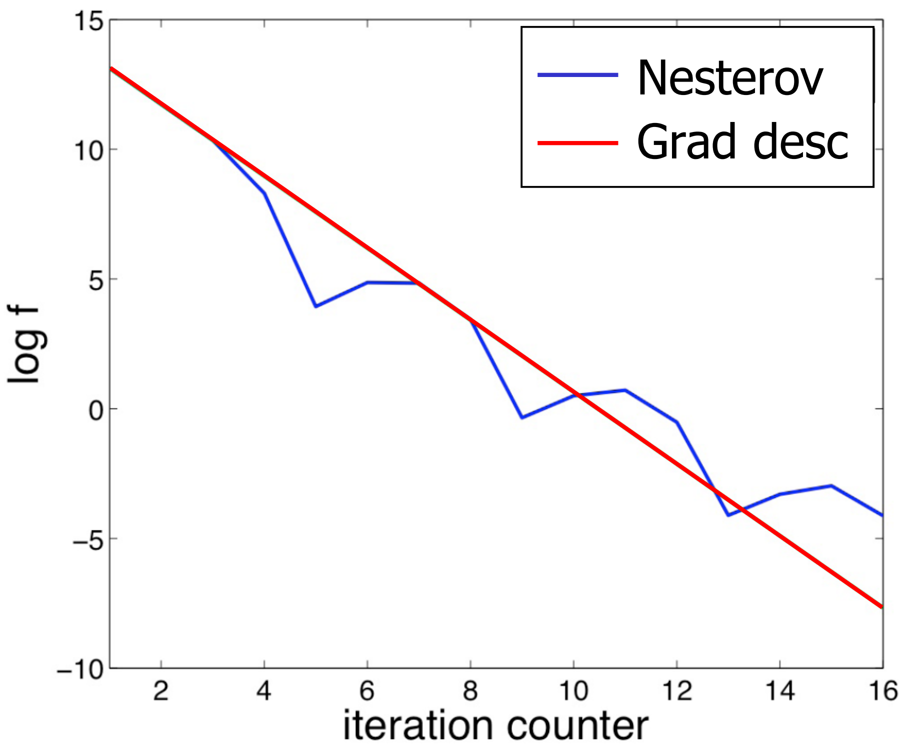
# Primitive Nesterov: Example

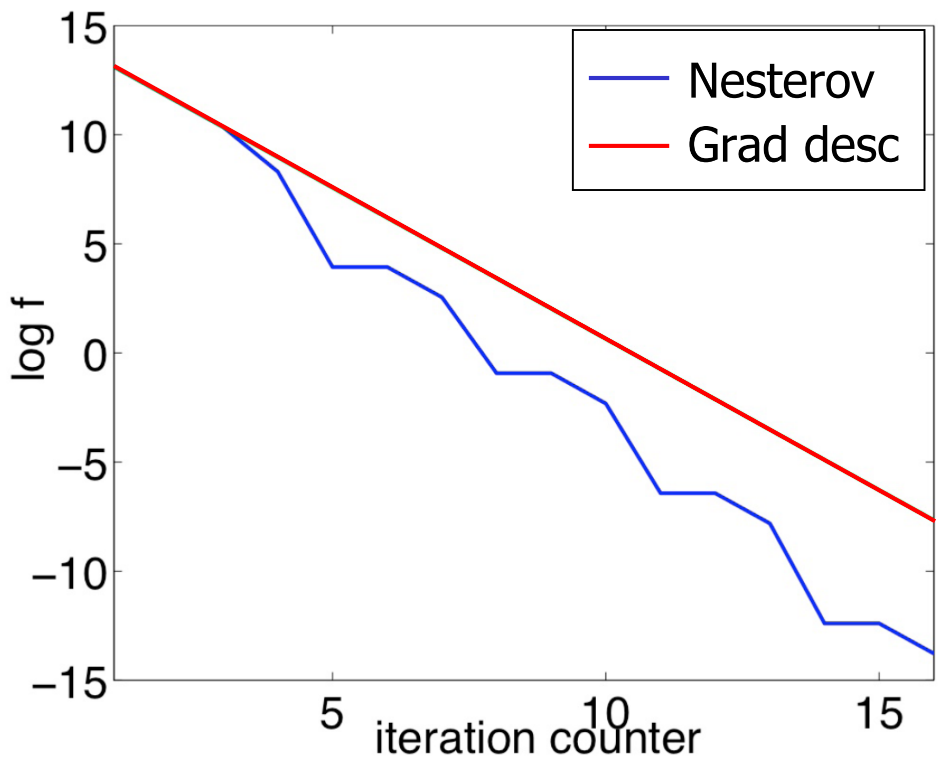
$$\min_{x,y} \frac{1}{2}x^2 + 2y^2$$

$$\mu = 1, \quad L = 4$$

$$\min_{x \geq 0, y \geq 0} \frac{1}{2}(x+y)^2$$

$$\mu = 0, \quad L = 2$$



# 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 $\|\mathbf{x}\|_1 = \sum_i x_i$
    - 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 $\|\mathbf{x}\|_1 = \sum_i x_i$
    - strongly convex wrt $L_1$ norm.

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

**Yes**

# 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

$$\min_{\mathbf{x}\in Q} d(\mathbf{x}) = 0 \qquad\qquad D := \max_{\mathbf{x}\in Q} d(\mathbf{x})$$

**Algorithm 1**: Nesterovs algorithm for non-Euclidean norm

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

1 Initialize: Set $\mathbf{x}^0$ to a random value in $Q$.

2 **for** $k = 0, 1, 2, \ldots$ **do**

3 Query the gradient of $f$ at point $\mathbf{x}^k$: $\nabla f(\mathbf{x}^k)$.

4 Find $\mathbf{y}^k \leftarrow \mathrm{argmin}_{\mathbf{x}\in Q} \left\langle \nabla f(\mathbf{x}^k), x - \mathbf{x}^k \right\rangle + \frac{1}{2} L \left\|\mathbf{x} - \mathbf{x}^k\right\|^2$ .

5 Find $\mathbf{z}^k \leftarrow \mathrm{argmin}_{\mathbf{x}\in Q} \frac{L}{\sigma} d(\mathbf{x}) + \sum_{i=0}^{k} \frac{i+1}{2} \left\langle \nabla f(\mathbf{x}^i), \mathbf{x} - \mathbf{x}^i \right\rangle$.

6 Update $\mathbf{x}^{k+1} \leftarrow \frac{2}{k+3} \mathbf{z}^k + \frac{k+1}{k+3} \mathbf{y}^k$.

I won't mention details

# 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

$$\min_{\mathbf{x}\in Q} d(\mathbf{x}) = 0 \qquad\qquad D := \max_{\mathbf{x}\in Q} d(\mathbf{x})$$

---

**Algorithm 1:** Nesterovs algorithm for non-Euclidean norm

---

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

Initialize: Set $\mathbf{x}^0$ to a random value in $Q$.

**for** $k = 0, 1, 2, \ldots$ **do**

Query the gradient of $f$ at point $\mathbf{x}^k$: $\nabla f(\mathbf{x}^k)$.

Find $\mathbf{y}^k \leftarrow \mathrm{argmin}_{\mathbf{x}\in Q} \left\langle \nabla f(\mathbf{x}^k), x - \mathbf{x}^k \right\rangle + \frac{1}{2} L \left\|\mathbf{x} - \mathbf{x}^k\right\|^2$ .

Find $\mathbf{z}^k \leftarrow \mathrm{argmin}_{\mathbf{x}\in Q} \frac{L}{\sigma} d(\mathbf{x}) + \sum_{i=0}^{k} \frac{i+1}{2} \left\langle \nabla f(\mathbf{x}^i), \mathbf{x} - \mathbf{x}^i \right\rangle$.

Update $\mathbf{x}^{k+1} \leftarrow \frac{2}{k+3}\mathbf{z}^k + \frac{k+1}{k+3}\mathbf{y}^k$.

---

I won't mention details

# Extension: Non-Euclidean norm

- Rate of convergence

$$f(\mathbf{y}_k) - f(\mathbf{x}^*) \leq \frac{4Ld(x^*)}{\sigma(k+1)(k+2)}$$

- Applications will be given later.

# Immediate application: Non-smooth functions

- Objective function not differentiable
  - Suppose it is the Fenchel dual of some function $f$

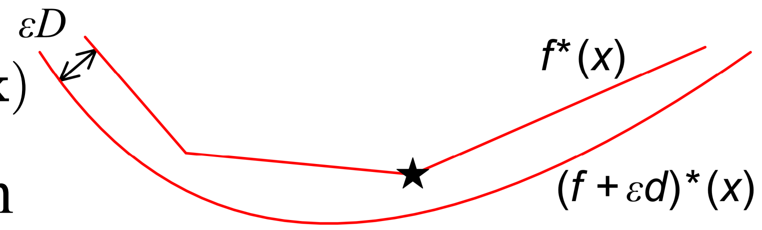
$$\min_{\mathbf{x}} f^{\star}(\mathbf{x}) \qquad \text{where } f \text{ is defined on } Q$$

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

$f + d$ is $\sigma$-strongly convex ⟹ $(f+d)^{\star}$ is $\frac{1}{\sigma}$-$l.c.g$

# Immediate application: Non-smooth functions

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




$$f^\star(\mathbf{x}) - \epsilon D \quad \leq \quad (f + \epsilon d)^\star(\mathbf{x}) \quad \leq \quad f^\star(\mathbf{x})$$

Proof

$$\max_{\mathbf{u}} \langle \mathbf{u}, \mathbf{x} \rangle - f(\mathbf{u}) - \epsilon D \leq \max_{\mathbf{u}} \langle \mathbf{u}, \mathbf{x} \rangle - f(\mathbf{u}) - \epsilon d(\mathbf{u}) \leq \max_{\mathbf{u}} \langle \mathbf{u}, \mathbf{x} \rangle - f(\mathbf{u}) - 0$$

$$\| \qquad\qquad \| \qquad\qquad \|$$

$$f^\star(\mathbf{x}) - \epsilon D \qquad (f + \epsilon d)^\star(\mathbf{x}) \qquad f^\star(\mathbf{x})$$

# Immediate application: Non-smooth functions

- $(f + \epsilon d)^\star(\mathbf{x})$ approximates $f^\star(\mathbf{x})$ well
  - If $d(u) \in [0, D]$ on $Q$, then $(f + \epsilon d)^\star(\mathbf{x}) - f^\star(\mathbf{x}) \in [-\epsilon D, 0]$
- Algorithm (given precision $\epsilon$)
  - Fix $\hat{\epsilon} = \frac{\epsilon}{2D}$
  - Optimize $(f + \hat{\epsilon} d)^\star(\mathbf{x})$ (*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

$$J(\mathbf{w}) = f(\mathbf{w}) + g^{\star}(A\mathbf{w})$$

where

$$f \text{ is convex on the region } E_1 \text{ with norm } \|\cdot\|_1$$

$$g \text{ is convex on the region } E_2 \text{ with norm } \|\cdot\|_2$$

- Very useful in machine learning
  - $A\mathbf{w}$ corresponds to linear model

# Composite optimization

- Example: binary SVM

$$J(\mathbf{w}) = \underbrace{\frac{\lambda}{2}\|\mathbf{w}\|^2}_{f(\mathbf{w})} + \underbrace{\min_{b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}[1 - y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b)]_+}_{g^\star(\mathbf{A}\mathbf{w})}$$

  - $A = -(y_1\mathbf{x}_1, \ldots, y_n\mathbf{x}_n)^\top$
  - $g^\star$ is the dual of $g(\boldsymbol{\alpha}) = -\sum_i \alpha_i$ over

$$Q_2 = \{\boldsymbol{\alpha} \in [0, n^{-1}]^n : \textstyle\sum_i y_i\alpha_i = 0\}$$

# Composite optimization 1: Smooth minimization

$$J(\mathbf{w}) = f(\mathbf{w}) + g^\star(A\mathbf{w})$$

- Let us only assume that

$$f \text{ is } M\text{-}l.c.g \text{ wrt } \|\cdot\|_1$$

- Smooth $g^\star$ into $(g+\mu d_2)^\star$ $(d_2 \text{ is } \sigma_2\text{-strongly convex wrt } \|\cdot\|_2)$

then $J_\mu(\mathbf{w}) = f(\mathbf{w}) + (g+\mu d_2)^\star(A\mathbf{w})$

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

Apply Nesterov on $J_\mu(\mathbf{w})$

# Composite optimization 1: Smooth minimization

- Rate of convergence
  - to find an $\epsilon$ accurate solution, it costs

$$4\|A\|_{1,2}\sqrt{\frac{D_1 D_2}{\sigma_1 \sigma_2}} \cdot \frac{1}{\epsilon} + \sqrt{\frac{M D_1}{\sigma_1 \epsilon}}$$

steps.

$$d_1 \text{ is } \sigma_1\text{-strongly convex wrt } \|\cdot\|_1$$

$$d_2 \text{ is } \sigma_2\text{-strongly convex wrt } \|\cdot\|_2$$

$$D_1 := \max_{\mathbf{w} \in E_1} d_1(\mathbf{w}) \qquad D_2 := \max_{\boldsymbol{\alpha} \in E_2} d_2(\boldsymbol{\alpha})$$

# Composite optimization 1: Smooth minimization

- Example: matrix game

$$\underset{\mathbf{w}\in\Delta_n}{\operatorname{argmin}} \underbrace{\langle \mathbf{c}, \mathbf{w}\rangle}_{f(\mathbf{w})} + \underbrace{\max_{\boldsymbol{\alpha}\in\Delta_m} \{\langle A\mathbf{w}, \boldsymbol{\alpha}\rangle + \langle \mathbf{b}, \boldsymbol{\alpha}\rangle\}}_{g^\star(A\mathbf{w})}$$

  - Use Euclidean distance

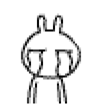
$$E_1 = \Delta_n \quad \|\mathbf{w}\|_1 = \left(\textstyle\sum_i w_i^2\right)^{1/2} \quad d_1(\mathbf{w}) = \tfrac{1}{2}\textstyle\sum_i (w_i - n^{-1})^2 \quad \sigma_1 = \sigma_2 = 1$$

$$E_2 = \Delta_m \quad \|\boldsymbol{\alpha}\|_2 = \left(\textstyle\sum_i \alpha_i^2\right)^{1/2} \quad d_2(\boldsymbol{\alpha}) = \tfrac{1}{2}\textstyle\sum_i (\alpha_i - m^{-1})^2 \quad D_1 < 1,\ D_2 < 1$$

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

$$f(\mathbf{w}_k) - f(\mathbf{w}^*) \le \frac{4\lambda_{\max}^{1/2}(A^\top A)}{k+1}$$

May scale with $O(nm)$



# Composite optimization 1: Smooth minimization

- Example: matrix game

$$\underset{\mathbf{w}\in\Delta_n}{\operatorname{argmin}} \underbrace{\langle \mathbf{c}, \mathbf{w}\rangle}_{f(\mathbf{w})} + \underbrace{\max_{\boldsymbol{\alpha}\in\Delta_m} \{\langle A\mathbf{w}, \boldsymbol{\alpha}\rangle + \langle \mathbf{b}, \boldsymbol{\alpha}\rangle\}}_{g^\star(A\mathbf{w})}$$

  - Use Entropy distance

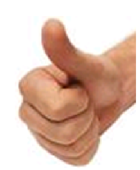
$$\sigma_1 = \sigma_2 = 1$$

$$E_1 = \Delta_n \quad \|\mathbf{w}\|_1 = \sum_i |w_i| \quad d_1(\mathbf{w}) = \ln n + \sum_i w_i \ln w_i \qquad D_1 = \ln n$$

$$E_2 = \Delta_m \quad \|\boldsymbol{\alpha}\|_2 = \sum_i |\alpha_i| \quad d_2(\boldsymbol{\alpha}) = \ln m + \sum_i \alpha_i \ln \alpha_i \qquad D_2 = \ln m$$

$$\|A\|_{1,2} = \max_{i,j} |A_{i,j}|$$

$$f(\mathbf{w}_k) - f(\mathbf{w}^*) \leq \frac{4\,(\ln n \ln m)^{\frac{1}{2}}}{k+1} \max_{i,j} \|A_{i,j}\|$$



# 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

$$J(\mathbf{w}) = f(\mathbf{w}) + g^{\star}(A\mathbf{w})$$

  - Utilizes the *adjoint* form

$$D(\boldsymbol{\alpha}) = -g(\boldsymbol{\alpha}) - f^{\star}(-A^{\top}\boldsymbol{\alpha})$$

  - Relations:

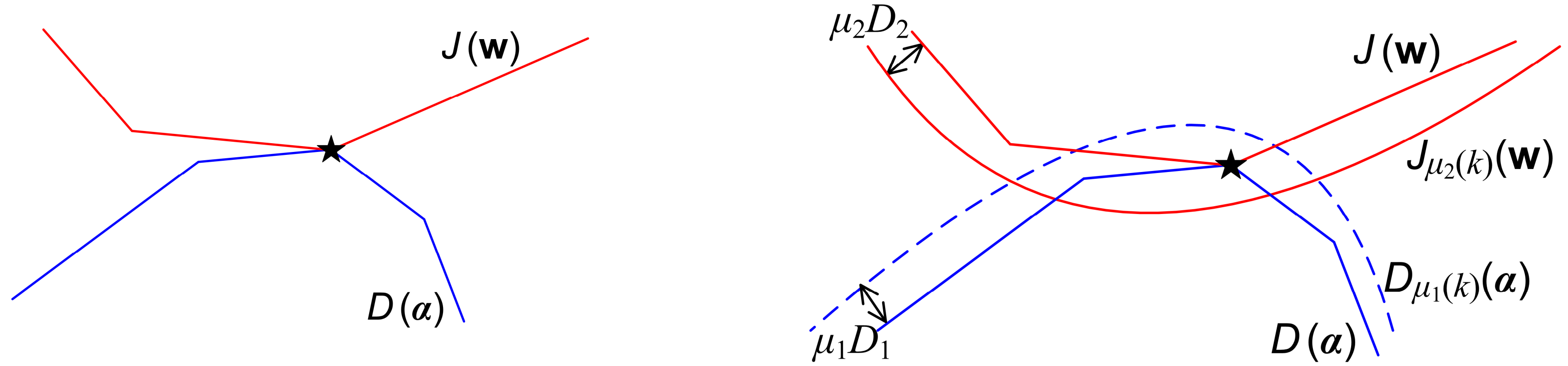
$\forall\ \mathbf{w}, \boldsymbol{\alpha} \quad J(\mathbf{w}) \geq D(\boldsymbol{\alpha})$ and $\inf_{\mathbf{w} \in E_1} J(\mathbf{w}) = \sup_{\boldsymbol{\alpha} \in E_2} D(\boldsymbol{\alpha})$

# 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 $f^\star$ and $g^\star$ by prox-functions $d_1, d_2$

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

$$D_{\mu_1}(\boldsymbol{\alpha}) = -g(\boldsymbol{\alpha}) - (f + \mu_1 d_1)^\star (-A^\top \boldsymbol{\alpha})$$

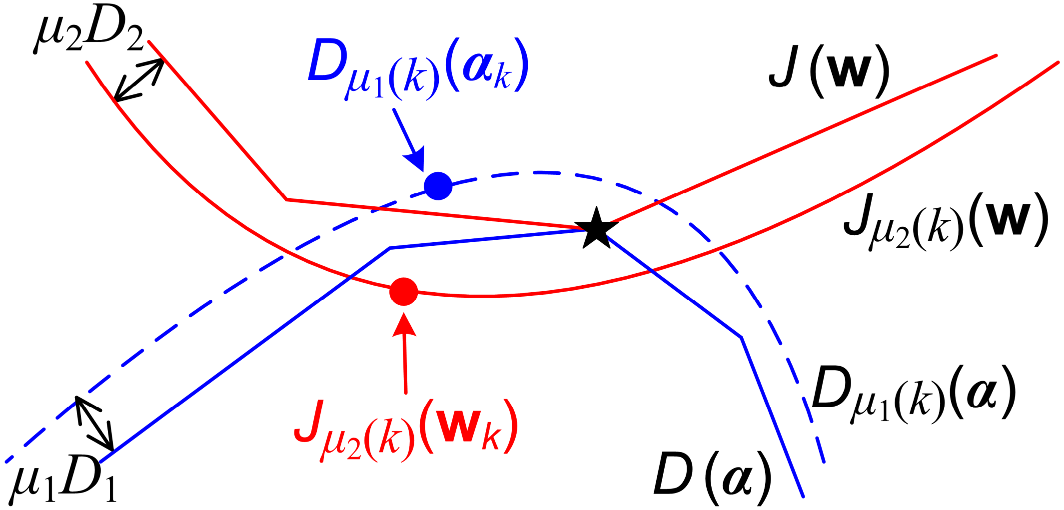


# Composite optimization 2: Excessive gap minimization

- Sketch of idea
  - Maintain two point sequences $\{\mathbf{w}_k\}$ and $\{\boldsymbol{\alpha}_k\}$ and two regularization sequences $\{\mu_1(k)\}$ and $\{\mu_2(k)\}$ s.t.

$$J_{\mu_2(k)}(\mathbf{w}_k) \leq D_{\mu_1(k)}(\boldsymbol{\alpha}_k)$$

$$\mu_1(k) \to 0$$
$$\mu_2(k) \to 0$$



# Composite optimization 2: Excessive gap minimization

$$J_{\mu_2(k)}(\mathbf{w}_k) \le D_{\mu_1(k)}(\boldsymbol{\alpha}_k)$$

- Challenge:
  - How to efficiently find the initial point $\mathbf{w}_1, \boldsymbol{\alpha}_1, \mu_1(1), \mu_2(1)$ that satisfy excessive gap condition.
  - Given $\mathbf{w}_k$, $\boldsymbol{\alpha}_k$, $\mu_1(k)$, $\mu_2(k)$, with new $\mu_1(k+1)$ and $\mu_2(k+1)$ how to efficiently find $\mathbf{w}_{k+1}$ and $\boldsymbol{\alpha}_{k+1}$.
  - 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:

$$J(\mathbf{w}_k) - D(\boldsymbol{\alpha}_k) \leq \frac{4\,\|A\|_{1,2}}{k+1}\sqrt{\frac{D_1 D_2}{\sigma_1 \sigma_2}}$$

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

$$J(\mathbf{w}_k) - D(\boldsymbol{\alpha}_k) \leq \frac{4D_2}{\sigma_2 k(k+1)}\left(\frac{\|A\|_{1,2}^2}{\sigma} + L_g\right)$$

# Composite optimization 2: Excessive gap minimization

- Example: binary SVM

$$J(\mathbf{w}) = \underbrace{\frac{\lambda}{2}\|\mathbf{w}\|^2}_{f(\mathbf{w})} + \underbrace{\min_{b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}[1 - y_i(\langle \mathbf{x}_i, \mathbf{w}\rangle + b)]_+}_{g^\star(\mathbf{Aw})}$$

  - $A = -(y_1\mathbf{x}_1, \ldots, y_n\mathbf{x}_n)^\top$
  - $g^\star$ is the dual of $g(\boldsymbol{\alpha}) = -\sum_i \alpha_i$ over

$$E_2 = \left\{\boldsymbol{\alpha} \in [0, n^{-1}]^n : \sum_i y_i\alpha_i = 0\right\}$$

  - Adjoint form $D(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2\lambda}\boldsymbol{\alpha}^\top A A^\top \boldsymbol{\alpha}$

# Composite optimization 2: Convergence rate for SVM

- Theorem: running on SVM for *k* iterations

$$J(\mathbf{w}_k) - D(\boldsymbol{\alpha}_k) \leq \frac{2L}{(k+1)(k+2)n}$$

  - $L = \lambda^{-1} \|A\|^2 = \lambda^{-1} \|(y_1\mathbf{x}_1, \ldots, y_n\mathbf{x}_n)\|^2 \leq \frac{nR^2}{\lambda} \quad (\|\mathbf{x}_i\| \leq R)$

- Final conclusion

$J(\mathbf{w}_k) - D(\boldsymbol{\alpha}_k) \leq \varepsilon$ as long as $k > O\left(\frac{R}{\sqrt{\lambda\,\varepsilon}}\right)$

# Composite optimization 2: Projection for SVM

- Efficient $O(n)$ time projection onto

$$E_2 = \left\{ \boldsymbol{\alpha} \in [0, n^{-1}]^n : \sum_i y_i \alpha_i = 0 \right\}$$

- Projection leads to a singly linear constrained $QP$

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} (\alpha_i - m_i)^2$$

$$s.t. \qquad l_i \le \alpha_i \le u_i \quad \forall i \in [n];$$

$$\sum_{i=1}^{n} \sigma_i \alpha_i = z.$$

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?