Outline - I

1. **Introduction**
   - Large-scale machine learning and optimization
   - Classes of functions (convex, smooth, etc.)
   - Traditional statistical analysis through Rademacher complexity

2. **Classical methods for convex optimization**
   - Smooth optimization (gradient descent, Newton method)
   - Non-smooth optimization (subgradient descent)
   - Proximal methods

3. **Non-smooth stochastic approximation**
   - Stochastic (sub)gradient and averaging
   - Non-asymptotic results and lower bounds
   - Strongly convex vs. non-strongly convex

Outline - II

4. **Classical stochastic approximation**
   - Asymptotic analysis
   - Robbins-Monro algorithm
   - Polyak-Rupert averaging

5. **Smooth stochastic approximation algorithms**
   - Non-asymptotic analysis for smooth functions
   - Logistic regression
   - Least-squares regression without decaying step-sizes

6. **Finite data sets**
   - Gradient methods with exponential convergence rates
   - Convex duality
   - (Dual) stochastic coordinate descent - Frank-Wolfe

Complexity results in convex optimization

- **Assumption**: g convex on $\mathbb{R}^d$

- **Classical generic algorithms**
  - Gradient descent and accelerated gradient descent
  - Newton method
  - Subgradient method and ellipsoid algorithm

Complexity results in convex optimization

- **Assumption**: g convex on $\mathbb{R}^d$
- **Classical generic algorithms**
  - Gradient descent and accelerated gradient descent
  - Newton method
  - Subgradient method and ellipsoid algorithm
- **Key additional properties of** g
  - Lipschitz continuity, smoothness or strong convexity
- **Key insight from Bottou and Bousquet (2008)**
  - In machine learning, no need to optimize below estimation error
- **Key references**: Nesterov (2004), Bubeck (2015)

Several criteria for characterizing convergence

- **Objective function values**

$$g(\theta) - \inf_{\eta \in \mathbb{R}^d} g(\eta)$$

  - Usually weaker condition

- **Iterates**

$$\inf_{\eta \in \arg\min g} \left\| \theta - \eta \right\|^2$$

  - Typically used for strongly-convex problems

- NB 1: relationships between the two types in several situations (see later)

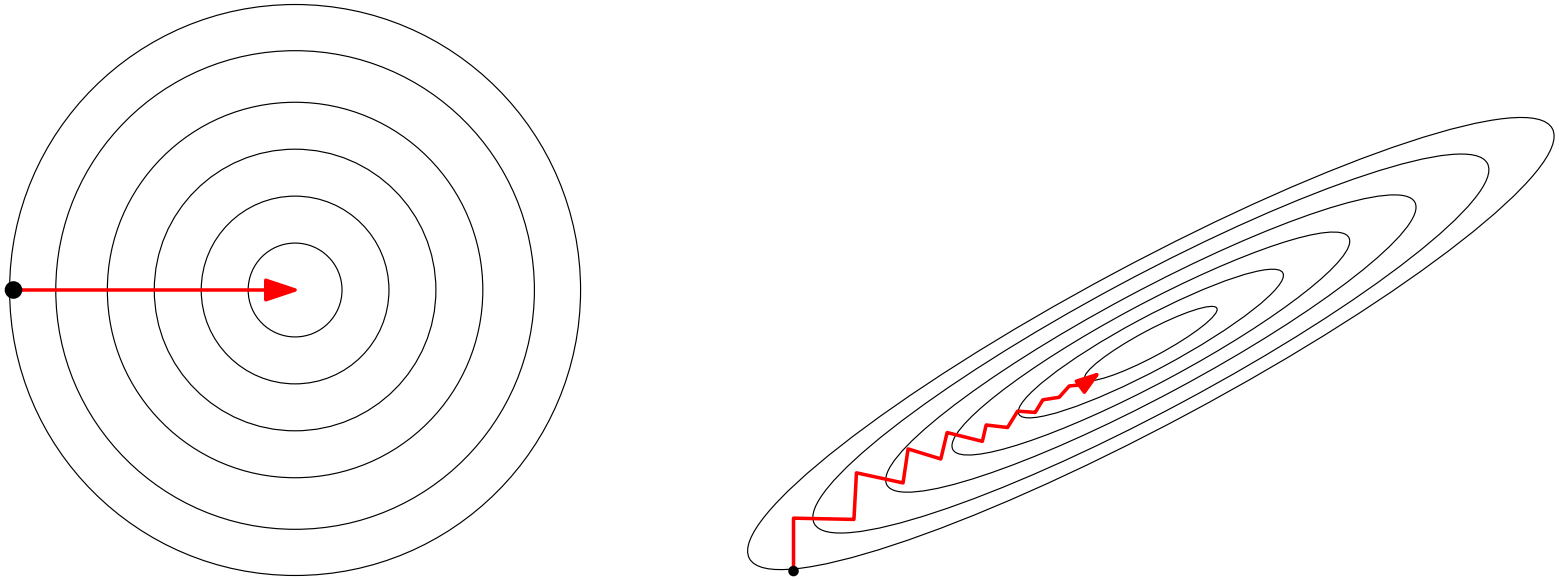
- NB 2: similarity with prediction vs. estimation in statistics

(smooth) gradient descent

- **Assumptions**
  - g convex with L-Lipschitz-continuous gradient (e.g., L-smooth)

- **Algorithm**:

$$\theta_t = \theta_{t-1} - \frac{1}{L} g'(\theta_{t-1})$$



(smooth) gradient descent - strong convexity

- **Assumptions**
  - g convex with L-Lipschitz-continuous gradient (e.g., L-smooth)
  - g μ-strongly convex
- **Algorithm**:

$$\theta_t = \theta_{t-1} - \frac{1}{L} g'(\theta_{t-1})$$

- **Bound**:

$$g(\theta_t) - g(\theta_*) \leqslant (1 - \mu/L)^t \big[g(\theta_0) - g(\theta_*) \big]$$

- Three-line proof
- **Line search, steepest descent or constant step-size**

(smooth) gradient descent - slow rate

- **Assumptions**
  - g convex with L-Lipschitz-continuous gradient (e.g., L-smooth)
  - Minimum attained at θ_*
- **Algorithm**:

$$\theta_t = \theta_{t-1} - \frac{1}{L} g'(\theta_{t-1})$$

- **Bound**:

$$g(\theta_t) - g(\theta_*) \leqslant \frac{2L\|\theta_0 - \theta_*\|^2}{t+4}$$

- Four-line proof
- **Adaptivity of gradient descent to problem difficulty**
- Not best possible convergence rates after $O(d)$ iterations

Gradient descent - Proof for quadratic functions

- Quadratic convex function: $g(\theta) = \frac{1}{2}\theta^\top H\theta - c^\top\theta$
  - μ and L are smallest largest eigenvalues of H
  - Global optimum $\theta_* = H^{-1}c$ (or $H^\dagger c$)

- Gradient descent:

$$\begin{aligned}\theta_t &= \theta_{t-1} - \frac{1}{L}(H\theta_{t-1} - c) = \theta_{t-1} - \frac{1}{L}(H\theta_{t-1} - H\theta_*) \\ \theta_t - \theta_* &= (I - \frac{1}{L}H)(\theta_{t-1} - \theta_*) = (I - \frac{1}{L}H)^t(\theta_0 - \theta_*)\end{aligned}$$

- **Strong convexity** $\mu > 0$: eigenvalues of $(I - \frac{1}{L}H)^t$ in $[0, (1 - \frac{\mu}{L})^t]$
  - Convergence of iterates: $\|\theta_t - \theta_*\|^2 \leqslant (1 - \mu/L)^{2t}\|\theta_0 - \theta_*\|^2$
  - Function values: $g(\theta_t) - g(\theta_*) \leqslant (1 - \mu/L)^{2t}\big[g(\theta_0) - g(\theta_*)\big]$

Gradient descent - Proof for quadratic functions

- Quadratic convex function: $g(\theta) = \frac{1}{2}\theta^\top H\theta - c^\top\theta$
  - μ and L are smallest largest eigenvalues of H
  - Global optimum $\theta_* = H^{-1}c$ (or $H^\dagger c$)

- Gradient descent:

$$\theta_t = \theta_{t-1} - \frac{1}{L}(H\theta_{t-1} - c) = \theta_{t-1} - \frac{1}{L}(H\theta_{t-1} - H\theta_*)$$
$$\theta_t - \theta_* = (I - \frac{1}{L}H)(\theta_{t-1} - \theta_*) = (I - \frac{1}{L}H)^t(\theta_0 - \theta_*)$$

- **Convexity** $\mu = 0$: eigenvalues of $(I - \frac{1}{L}H)^t$ in $[0, 1]$
  - No convergence of iterates: $\|\theta_t - \theta_*\|^2 \leqslant \|\theta_0 - \theta_*\|^2$
  - Function values: $g(\theta_t) - g(\theta_*) \leqslant \max_{v\in[0,L]} v(1-v/L)^{2t}\|\theta_0 - \theta_*\|^2$

$$g(\theta_t) - g(\theta_*) \leqslant \frac{L}{t}\|\theta_0 - \theta_*\|^2$$

Properties of smooth convex functions

- Let $g : \mathbb{R}^d \to \mathbb{R}$ a convex L-smooth function. Then for all $\theta, \eta \in \mathbb{R}^d$:
  - Definition: $\|g'(\theta) - g'(\eta)\| \leqslant L\|\theta - \eta\|$
  - If twice differentiable: $0 \preccurlyeq g''(\theta) \preccurlyeq LI$
- Quadratic upper-bound: $0 \leqslant g(\theta) - g(\eta) - g'(\eta)^\top(\theta - \eta) \leqslant \frac{L}{2}\|\theta - \eta\|^2$
  - Taylor expansion with integral remainder
- Co-coercivity: $\frac{1}{L}\|g'(\theta) - g'(\eta)\|^2 \leqslant \big[g'(\theta) - g'(\eta)\big]^\top(\theta - \eta)$
- If g is μ-strongly convex (no need for smoothness), then

$$g(\theta) \leqslant g(\eta) + g'(\eta)^\top(\theta - \eta) + \frac{1}{2\mu}\|g'(\theta) - g'(\eta)\|^2$$

- "Distance" to optimum: $g(\theta) - g(\theta_*) \leqslant g'(\theta)^\top(\theta - \theta_*)$

Proof of co-coercivity

- Quadratic upper-bound: $0 \leqslant g(\theta) - g(\eta) - g'(\eta)^\top(\theta - \eta) \leqslant \frac{L}{2}\|\theta - \eta\|^2$
  - Taylor expansion with integral remainder
- Lower bound: $g(\theta) \geqslant g(\eta) + g'(\eta)^\top(\theta - \eta) + \frac{1}{2L}\|g'(\theta) - g'(\eta)\|^2$
  - Define $h(\theta) = g(\theta) - \theta^\top g'(\eta)$, convex with global minimum at η
  - $h(\eta) \leqslant h(\theta - \frac{1}{L}h'(\theta)) \leqslant h(\theta) + h'(\theta)^\top(-\frac{1}{L}h'(\theta))) + \frac{L}{2}\|-\frac{1}{L}h'(\theta))\|^2$, which is thus less than $h(\theta) - \frac{1}{2L}\|h'(\theta)\|^2$
  - Thus $g(\eta) - \eta^\top g'(\eta) \leqslant g(\theta) - \theta^\top g'(\eta) - \frac{1}{2L}\|g'(\theta) - g'(\eta)\|^2$
- Proof of co-coercivity
  - Apply lower bound twice for (η, θ) and (θ, η), and sum to get $0 \geqslant [g'(\eta) - g'(\theta)]^\top(\theta - \eta) + \frac{1}{L}\|g'(\theta) - g'(\eta)\|^2$
- NB: simple proofs with second-order derivatives

Proof of $g(\theta) \leqslant g(\eta) + g'(\eta)^\top(\theta - \eta) + \frac{1}{2\mu}\|g'(\theta) - g'(\eta)\|^2$

- Define $h(\theta) = g(\theta) - \theta^\top g'(\eta)$, convex with global minimum at η
- $h(\eta) = \min_\theta h(\theta) \geqslant \min_\zeta h(\theta) + h'(\theta)^\top(\zeta - \theta) + \frac{\mu}{2}\|\zeta - \theta\|^2$, which is attained for $\zeta - \theta = -\frac{1}{\mu}h'(\theta)$
  - This leads to $h(\eta) \geqslant h(\theta) - \frac{1}{2\mu}\|h'(\theta)\|^2$
  - Hence, $g(\eta) - \eta^\top g'(\eta) \geqslant g(\theta) - \theta^\top g'(\eta) - \frac{1}{2\mu}\|g'(\eta) -' (\theta)\|^2$
  - NB: no need for smooothness
- NB: simple proofs with second-order derivatives
- With $\eta = \theta_*$ global minimizer, another "distance" to optimum

$$g(\theta) - g(\theta_*) \leqslant \frac{1}{2\mu}\|g'(\theta)\|^2 \quad \text{"Polyak-Lojasiewicz"}$$

Convergence proofs through Lyapunov functions

- Given sequence of iterates (θ_t), find a function $V \geqslant 0$ such that

$$V(\theta_t) \leqslant (1-\alpha)V(\theta_{t-1})$$

  - Then $V(\theta_t) \leqslant (1-\alpha)^t V(\alpha_0)$

- Many variations
  - Time-dependence: $V_t(\theta_t) \leqslant (1-\alpha_t)V_{t-1}(\theta_{t-1})$
  - Weak decrease: $V(\theta_t) \leqslant V(\theta_{t-1}) - U(\theta_t)$
    Then $U(\theta_t) \leqslant V(\theta_{t-1}) - V(\theta_t)$ and $\frac{1}{T}\sum_{t=1}^{T} U(\theta_t) \leqslant \frac{V(\theta_0)}{T}$

- Noise term: $V(\theta_t) \leqslant V(\theta_{t-1}) - U(\theta_t) + M(\theta_{t-1})$

- Classical candidates: $\|\theta - \theta_*\|_2^2$ and $g(\theta) - g(\theta_*)$

Convergence proof - gradient descent smooth strongly convex functions

- Iteration: $\theta_t = \theta_{t-1} - \gamma g'(\theta_{t-1})$ with $\gamma = 1/L$

$$\begin{aligned} g(\theta_t) &= g\big[\theta_{t-1} - \gamma g'(\theta_{t-1})\big] \leqslant g(\theta_{t-1}) + g'(\theta_{t-1})^\top\big[-\gamma g'(\theta_{t-1})\big] + \frac{L}{2}\|-\gamma g'(\theta_{t-1})\|^2 \\ &= g(\theta_{t-1}) - \gamma(1-\gamma L/2)\|g'(\theta_{t-1})\|^2 \\ &= g(\theta_{t-1}) - \frac{1}{2L}\|g'(\theta_{t-1})\|^2 \text{ if } \gamma = 1/L, \\ &\leqslant g(\theta_{t-1}) - \frac{\mu}{L}\big[g(\theta_{t-1}) - g(\theta_*)\big] \text{ using strongly-convex "distance" to optimum} \end{aligned}$$

Thus, $g(\theta_t) - g(\theta_*) \leqslant (1-\mu/L)\big[g(\theta_{t-1}) - g(\theta_*)\big] \leqslant (1-\mu/L)^t\big[g(\theta_0) - g(\theta_*)\big]$

- May also get (Nesterov, 2004): $\|\theta_t - \theta_*\|^2 \leqslant \big(1 - \frac{2\gamma\mu L}{\mu+L}\big)^t\|\theta_0 - \theta_*\|^2$ as soon as $\gamma \leqslant \frac{2}{\mu+L}$

Convergence proof - gradient descent smooth convex functions - I

- Iteration: $\theta_t = \theta_{t-1} - \gamma g'(\theta_{t-1})$ with $\gamma = 1/L$

$$
\begin{aligned}
\|\theta_t - \theta_*\|^2 &= \|\theta_{t-1} - \theta_* - \gamma g'(\theta_{t-1})\|^2 \\
&= \|\theta_{t-1} - \theta_*\|^2 + \gamma^2 \|g'(\theta_{t-1})\|^2 - 2\gamma (\theta_{t-1} - \theta_*)^\top g'(\theta_{t-1}) \\
&\leqslant \|\theta_{t-1} - \theta_*\|^2 + \gamma^2 \|g'(\theta_{t-1})\|^2 - 2\frac{\gamma}{L} \|g'(\theta_{t-1})\|^2 \text{ using co-coercivity} \\
&= \|\theta_{t-1} - \theta_*\|^2 - \gamma(2/L - \gamma)\|g'(\theta_{t-1})\|^2 \leqslant \|\theta_{t-1} - \theta_*\|^2 \text{ if } \gamma \leqslant 2/L \\
&\leqslant \|\theta_0 - \theta_*\|^2 : \text{ bounded iterates} \\
g(\theta_t) &\leqslant g(\theta_{t-1}) - \frac{1}{2L} \|g'(\theta_{t-1})\|^2 \text{ (see previous slide)} \\
g(\theta_{t-1}) - g(\theta_*) &\leqslant g'(\theta_{t-1})^\top (\theta_{t-1} - \theta_*) \leqslant \|g'(\theta_{t-1})\| \cdot \|\theta_{t-1} - \theta_*\| \text{ (Cauchy-Schwarz)} \\
g(\theta_t) - g(\theta_*) &\leqslant g(\theta_{t-1}) - g(\theta_*) - \frac{1}{2L\|\theta_0 - \theta_*\|^2} \big[g(\theta_{t-1}) - g(\theta_*)\big]^2
\end{aligned}
$$

Convergence proof - gradient descent smooth convex functions - II

- Iteration: $\theta_t = \theta_{t-1} - \gamma g'(\theta_{t-1})$ with $\gamma = 1/L$

$$
\begin{aligned}
g(\theta_t) - g(\theta_*) &\leqslant g(\theta_{t-1}) - g(\theta_*) - \frac{1}{2L\|\theta_0 - \theta_*\|^2}\big[g(\theta_{t-1}) - g(\theta_*)\big]^2 \\
\text{of the form } \Delta_k &\leqslant \Delta_{k-1} - \alpha\Delta_{k-1}^2 \text{ with } 0 \leqslant \Delta_k = g(\theta_k) - g(\theta_*) \leqslant \frac{L}{2}\|\theta_k - \theta_*\|^2 \\
\frac{1}{\Delta_{k-1}} &\leqslant \frac{1}{\Delta_k} - \alpha\frac{\Delta_{k-1}}{\Delta_k} \text{ by dividing by } \Delta_k\Delta_{k-1} \\
\frac{1}{\Delta_{k-1}} &\leqslant \frac{1}{\Delta_k} - \alpha \text{ because } (\Delta_k) \text{ is non-increasing} \\
\frac{1}{\Delta_0} &\leqslant \frac{1}{\Delta_t} - \alpha t \text{ by summing from } k = 1 \text{ to } t \\
\Delta_t &\leqslant \frac{\Delta_0}{1 + \alpha t \Delta_0} \text{ by inverting} \\
&\leqslant \frac{2L\|\theta_0 - \theta_*\|^2}{t+4} \text{ since } \Delta_0 \leqslant \frac{L}{2}\|\theta_k - \theta_*\|^2 \text{ and } \alpha = \frac{1}{2L\|\theta_0 - \theta_*\|^2}
\end{aligned}
$$

Limits on convergence rate of first-order methods

- **First-order method**: any iterative algorithm that selects θ_t in $\theta_0 + \operatorname{span}(g'(\theta_0), \ldots, g'(\theta_{t-1}))$
- **Problem class**: convex L-smooth functions with a global minimizer θ_*
- **Theorem**: for every integer $t \leqslant (d-1)/2$ and every θ_0, there exist functions in the problem class such that for any first-order method,

$$g(\theta_t) - g(\theta_*) \geqslant \frac{3}{32} \frac{L \|\theta_0 - \theta_*\|^2}{(t+1)^2}$$

  - $O(1/t)$ rate for gradient method may not be optimal!

Limits on convergence rate of first-order methods
Proof sketch

- Define quadratic function

$$g_t(\theta) = \frac{L}{8}\left[(\theta^1)^2 + \sum_{i=1}^{t-1}(\theta^i - \theta^{i+1})^2 + (\theta^t)^2 - 2\theta^1\right]$$

  - Fact 1: g_t is L-smooth
  - Fact 2: minimizer supported by first t coordinates (closed form)
  - Fact 3: any first-order method starting from zero will be supported in the first k coordinates after iteration k
  - Fact 4: the minimum over this support in $\{1, \ldots, k\}$ may be computed in closed form

- Given iteration k, take $g = g_{2k+1}$ and compute lower-bound on $\frac{g(\theta_k) - g(\theta_*)}{\|\theta_0 - \theta_*\|^2}$

Accelerated gradient methods (Nesterov, 1983)

- **Assumptions**
  - g convex with L-Lipschitz-cont. gradient , min. attained at θ_*
- **Algorithm**:

$$\begin{aligned} \theta_t &= \eta_{t-1} - \frac{1}{L} g'(\eta_{t-1}) \\ \eta_t &= \theta_t + \frac{t-1}{t+2}(\theta_t - \theta_{t-1}) \end{aligned}$$

- **Bound**:

$$g(\theta_t) - g(\theta_*) \leqslant \frac{2L\|\theta_0 - \theta_*\|^2}{(t+1)^2}$$

- Ten-line proof (see, e.g., Schmidt, Le Roux, and Bach, 2011)
- Not improvable
- Extension to strongly-convex functions

Accelerated gradient methods - strong convexity

- **Assumptions**
  - g convex with L-Lipschitz-cont. gradient , min. attained at θ_*
  - g μ-strongly convex

- **Algorithm**:

$$\begin{aligned}\theta_t &= \eta_{t-1} - \frac{1}{L} g'(\eta_{t-1}) \\ \eta_t &= \theta_t + \frac{1-\sqrt{\mu/L}}{1+\sqrt{\mu/L}}(\theta_t - \theta_{t-1})\end{aligned}$$

- **Bound**: $g(\theta_t) - g(\theta_*) \leqslant L\|\theta_0 - \theta_*\|^2(1-\sqrt{\mu/L})^t$
  - Ten-line proof (see, e.g., Schmidt, Le Roux, and Bach, 2011)
  - Not improvable
  - Relationship with conjugate gradient for quadratic functions

Optimization for sparsity-inducing norms (see Bach, Jenatton, Mairal, and Obozinski, 2012b)

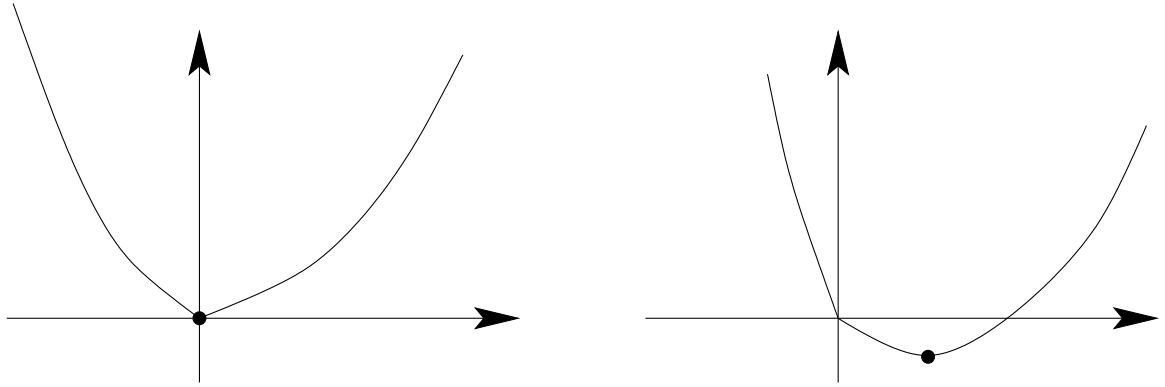
- Gradient descent as a **proximal method** (differentiable functions)
  - $\theta_{t+1} = \arg\min_{\theta \in \mathbb{R}^d} f(\theta_t) + (\theta - \theta_t)^\top \nabla f(\theta_t) + \frac{L}{2} \|\theta - \theta_t\|_2^2$
  - $\theta_{t+1} = \theta_t - \frac{1}{L} \nabla f(\theta_t)$

Optimization for sparsity-inducing norms (see Bach, Jenatton, Mairal, and Obozinski, 2012b)

- Gradient descent as a **proximal method** (differentiable functions)
  - $\theta_{t+1} = \arg\min_{\theta \in \mathbb{R}^d} f(\theta_t) + (\theta - \theta_t)^\top \nabla f(\theta_t) + \frac{L}{2}\|\theta - \theta_t\|_2^2$
  - $\theta_{t+1} = \theta_t - \frac{1}{L}\nabla f(\theta_t)$
- Problems of the form: $\boxed{\min_{\theta \in \mathbb{R}^d} f(\theta) + \mu\Omega(\theta)}$
  - $\theta_{t+1} = \arg\min_{\theta \in \mathbb{R}^d} f(\theta_t) + (\theta - \theta_t)^\top \nabla f(\theta_t) + \mu\Omega(\theta) + \frac{L}{2}\|\theta - \theta_t\|_2^2$
  - $\Omega(\theta) = \|\theta\|_1 \Rightarrow$ **Thresholded gradient descent**
- Similar convergence rates than smooth optimization
  - Acceleration methods (Nesterov, 2007; Beck and Teboulle, 2009)

Soft-thresholding for the ℓ_1-norm

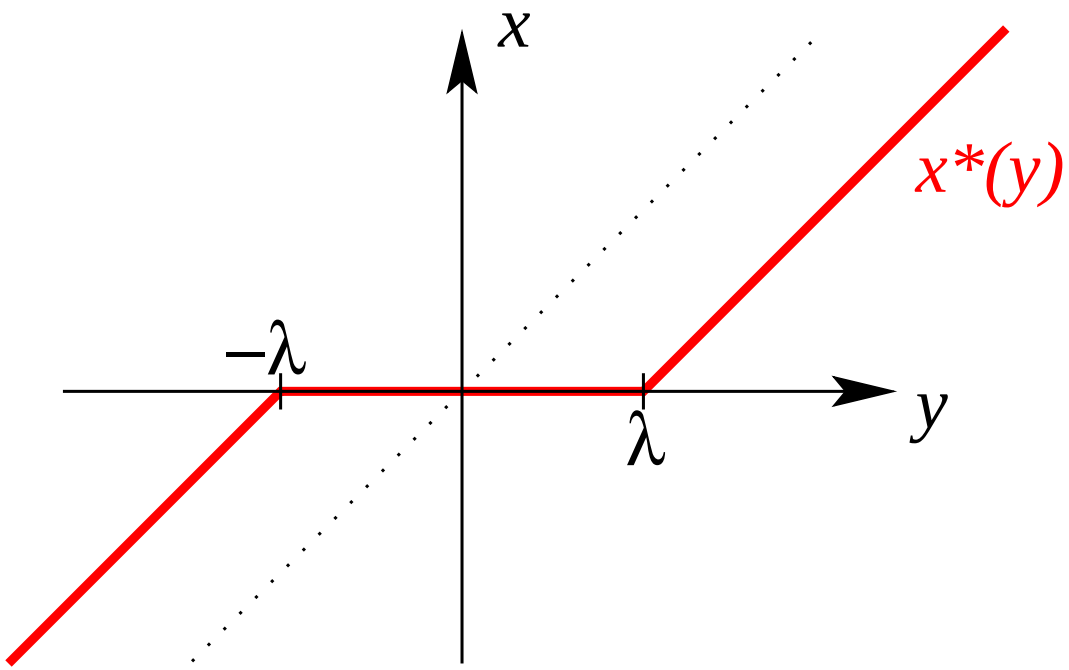
- **Example 1**: quadratic problem in 1D, i.e. $\boxed{\min_{x\in\mathbb{R}} \frac{1}{2}x^2 - xy + \lambda|x|}$
- Piecewise quadratic function with a kink at zero
  - Derivative at $0+$: $g_+ = \lambda - y$ and $0-$: $g_- = -\lambda - y$



  - $x = 0$ is the solution iff $g_+ \geqslant 0$ and $g_- \leqslant 0$ (i.e., $|y| \leqslant \lambda$)
  - $x \geqslant 0$ is the solution iff $g_+ \leqslant 0$ (i.e., $y \geqslant \lambda$) $\Rightarrow x^* = y - \lambda$
  - $x \leqslant 0$ is the solution iff $g_- \geqslant 0$ (i.e., $y \leqslant -\lambda$) $\Rightarrow x^* = y + \lambda$
- Solution $\boxed{x^* = \mathrm{sign}(y)(|y| - \lambda)_+}$ = soft thresholding

Soft-thresholding for the ℓ_1-norm

- **Example 1**: quadratic problem in 1D, i.e. $\boxed{\min_{x\in\mathbb{R}} \frac{1}{2}x^2 - xy + \lambda|x|}$
- Piecewise quadratic function with a kink at zero
- Solution $\boxed{x^* = \mathrm{sign}(y)(|y| - \lambda)_+}$ = soft thresholding



Projected gradient descent

- Problems of the form: $\boxed{\min_{\theta \in \mathcal{K}} f(\theta)}$
  - $\theta_{t+1} = \arg\min_{\theta \in \mathcal{K}} f(\theta_t) + (\theta - \theta_t)^\top \nabla f(\theta_t) + \frac{L}{2} \|\theta - \theta_t\|_2^2$
  - $\theta_{t+1} = \arg\min_{\theta \in \mathcal{K}} \frac{1}{2} \left\| \theta - \left(\theta_t - \frac{1}{L} \nabla f(\theta_t) \right) \right\|_2^2$
  - Projected gradient descent
- Similar convergence rates than smooth optimization
  - Acceleration methods (Nesterov, 2007; Beck and Teboulle, 2009)

Newton method

- Given θ_{t-1}, minimize second-order Taylor expansion

$$\tilde{g}(\theta) = g(\theta_{t-1}) + g'(\theta_{t-1})^\top(\theta - \theta_{t-1}) + \frac{1}{2}(\theta - \theta_{t-1})^\top g''(\theta_{t-1})^\top(\theta - \theta_{t-1})$$

- **Expensive Iteration**: $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1} g'(\theta_{t-1})$
  - Running-time complexity: $O(d^3)$ in general

- **Quadratic convergence**: If $\|\theta_{t-1} - \theta_*\|$ small enough, for some constant C, we have

$$(C\|\theta_t - \theta_*\|) = (C\|\theta_{t-1} - \theta_*\|)^2$$

  - See Boyd and Vandenberghe (2003)

Summary: minimizing smooth convex functions

- **Assumption**: g convex
- **Gradient descent**: $\theta_t = \theta_{t-1} - \gamma_t \, g'(\theta_{t-1})$
  - $O(1/t)$ convergence rate for smooth convex functions
  - $O(e^{-t\mu/L})$ convergence rate for strongly smooth convex functions
  - Optimal rates $O(1/t^2)$ and $O(e^{-t\sqrt{\mu/L}})$
- **Newton method**: $\theta_t = \theta_{t-1} - f''(\theta_{t-1})^{-1} f'(\theta_{t-1})$
  - $O\big(e^{-\rho 2^t}\big)$ convergence rate

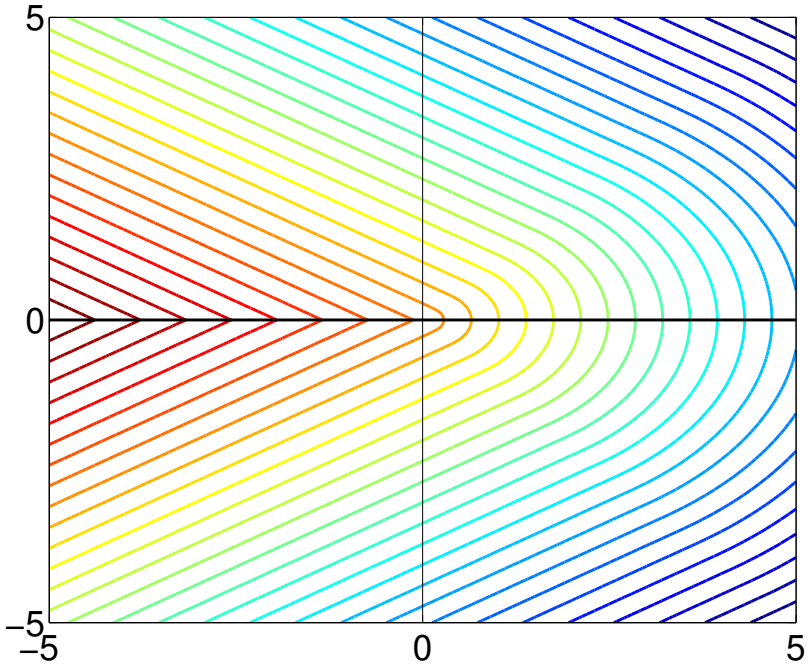
Summary: minimizing smooth convex functions

- **Assumption**: g convex
- **Gradient descent**: $\theta_t = \theta_{t-1} - \gamma_t \, g'(\theta_{t-1})$
  - $O(1/t)$ convergence rate for smooth convex functions
  - $O(e^{-t\mu/L})$ convergence rate for strongly smooth convex functions
  - Optimal rates $O(1/t^2)$ and $O(e^{-t\sqrt{\mu/L}})$
- **Newton method**: $\theta_t = \theta_{t-1} - f''(\theta_{t-1})^{-1} f'(\theta_{t-1})$
  - $O\big(e^{-\rho 2^t}\big)$ convergence rate
- **From smooth to non-smooth**
  - Subgradient method and ellipsoid

Counter-example (Bertsekas, 1999)
Steepest descent for nonsmooth objectives

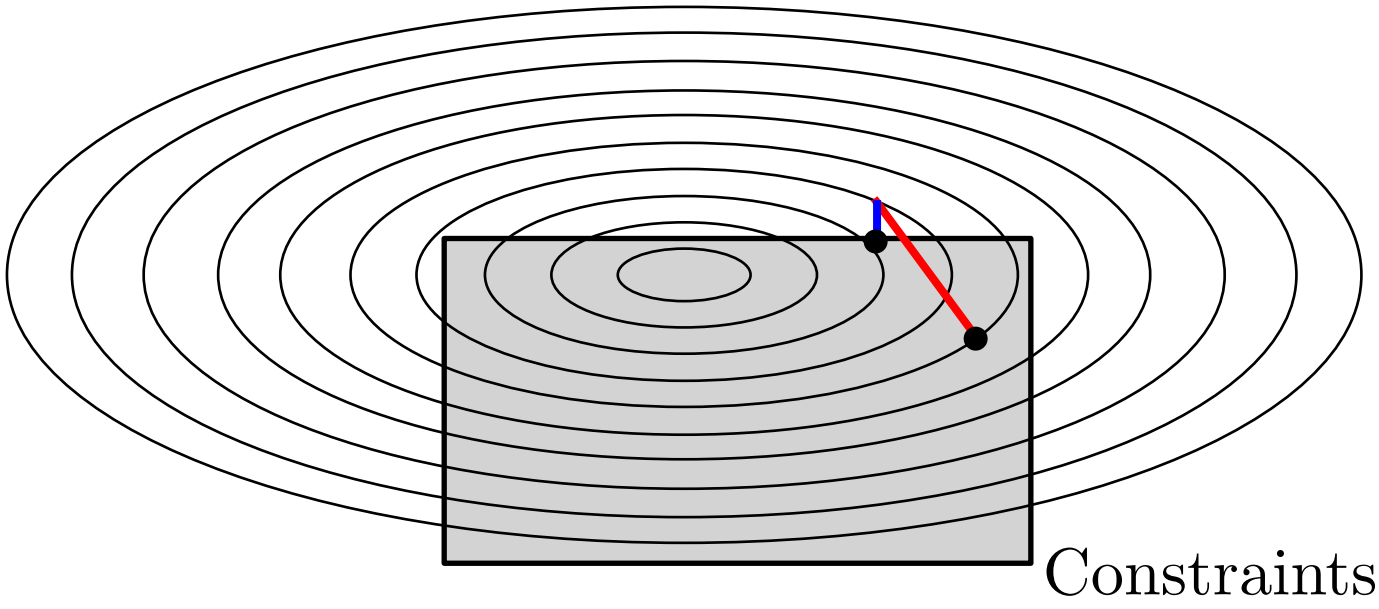
- $g(\theta_1, \theta_2) = \begin{cases} -5(9\theta_1^2 + 16\theta_2^2)^{1/2} \text{ if } \theta_1 > |\theta_2| \\ -(9\theta_1 + 16|\theta_2|)^{1/2} \text{ if } \theta_1 \leqslant |\theta_2| \end{cases}$

- Steepest descent starting from any θ such that $\theta_1 > |\theta_2| > (9/16)^2|\theta_1|$



Subgradient method/"descent" (Shor et al., 1985)

- **Assumptions**
  - g convex and B-Lipschitz-continuous on $\{\|\theta\|_2 \leqslant D\}$
- **Algorithm**: $\theta_t = \Pi_D\left(\theta_{t-1} - \dfrac{2D}{B\sqrt{t}} g'(\theta_{t-1})\right)$
  - Π_D : orthogonal projection onto $\{\|\theta\|_2 \leqslant D\}$



Subgradient method/"descent" (Shor et al., 1985)

- **Assumptions**
  - g convex and B-Lipschitz-continuous on $\{\|\theta\|_2 \leqslant D\}$
- **Algorithm**: $\theta_t = \Pi_D\left(\theta_{t-1} - \frac{2D}{B\sqrt{t}} g'(\theta_{t-1})\right)$
  - Π_D : orthogonal projection onto $\{\|\theta\|_2 \leqslant D\}$
- **Bound**:

$$g\left(\frac{1}{t}\sum_{k=0}^{t-1}\theta_k\right) - g(\theta_*) \leqslant \frac{2DB}{\sqrt{t}}$$

- Three-line proof
- Best possible convergence rate after $O(d)$ iterations (Bubeck, 2015)

Subgradient method/"descent" - proof - I

- Iteration: $\theta_t = \Pi_D(\theta_{t-1} - \gamma_t g'(\theta_{t-1}))$ with $\gamma_t = \frac{2D}{B\sqrt{t}}$
- Assumption: $\|g'(\theta)\|_2 \leqslant B$ and $\|\theta\|_2 \leqslant D$

$$
\begin{aligned}
\|\theta_t - \theta_*\|_2^2 \leqslant & \quad \|\theta_{t-1} - \theta_* - \gamma_t g'(\theta_{t-1})\|_2^2 \text{ by contractivity of projections} \\
\leqslant & \quad \|\theta_{t-1} - \theta_*\|_2^2 + B^2\gamma_t^2 - 2\gamma_t(\theta_{t-1} - \theta_*)^\top g'(\theta_{t-1}) \text{ because } \|g'(\theta_{t-1})\|_2 \leqslant B \\
\leqslant & \quad \|\theta_{t-1} - \theta_*\|_2^2 + B^2\gamma_t^2 - 2\gamma_t\big[g(\theta_{t-1}) - g(\theta_*)\big] \text{ (property of subgradients)}
\end{aligned}
$$

- leading to

$$
g(\theta_{t-1}) - g(\theta_*) \leqslant \frac{B^2\gamma_t}{2} + \frac{1}{2\gamma_t}\big[\|\theta_{t-1} - \theta_*\|_2^2 - \|\theta_t - \theta_*\|_2^2\big]
$$

Subgradient method/"descent" - proof - II

- Starting from $g(\theta_{t-1}) - g(\theta_*) \leqslant \frac{B^2\gamma_t}{2} + \frac{1}{2\gamma_t}\left[\|\theta_{t-1} - \theta_*\|_2^2 - \|\theta_t - \theta_*\|_2^2\right]$
- Constant step-size $\gamma_t = \gamma$

$$\begin{aligned}\sum_{u=1}^{t}\left[g(\theta_{u-1}) - g(\theta_*)\right] \leqslant & \sum_{u=1}^{t}\frac{B^2\gamma}{2} + \sum_{u=1}^{t}\frac{1}{2\gamma}\left[\|\theta_{u-1} - \theta_*\|_2^2 - \|\theta_u - \theta_*\|_2^2\right] \\ \leqslant & \; t\frac{B^2\gamma}{2} + \frac{1}{2\gamma}\|\theta_0 - \theta_*\|_2^2 \leqslant t\frac{B^2\gamma}{2} + \frac{2}{\gamma}D^2\end{aligned}$$

- Optimized step-size $\gamma_T = \frac{2D}{B\sqrt{T}}$ depends on "horizon" T
  - Leads to bound of $2DB\sqrt{T}$
- Using convexity: $g\left(\frac{1}{T}\sum_{k=0}^{T-1}\theta_k\right) - g(\theta_*) \leqslant \frac{1}{T}\sum_{k=0}^{T-1} g(\theta_k) - g(\theta_*) \leqslant \frac{2DB}{\sqrt{T}}$

Subgradient method/"descent" - proof - III

- Starting from $g(\theta_{t-1}) - g(\theta_*) \leqslant \frac{B^2\gamma_t}{2} + \frac{1}{2\gamma_t}\left[\|\theta_{t-1} - \theta_*\|_2^2 - \|\theta_t - \theta_*\|_2^2\right]$
- Decreasing step-size

$$\begin{aligned}
\sum_{u=1}^{t}\left[g(\theta_{u-1}) - g(\theta_*)\right] \leqslant & \sum_{u=1}^{t}\frac{B^2\gamma_u}{2} + \sum_{u=1}^{t}\frac{1}{2\gamma_u}\left[\|\theta_{u-1} - \theta_*\|_2^2 - \|\theta_u - \theta_*\|_2^2\right] \\
= & \sum_{u=1}^{t}\frac{B^2\gamma_u}{2} + \sum_{u=1}^{t-1}\|\theta_u - \theta_*\|_2^2\left(\frac{1}{2\gamma_{u+1}} - \frac{1}{2\gamma_u}\right) + \frac{\|\theta_0 - \theta_*\|_2^2}{2\gamma_1} - \frac{\|\theta_t - \theta_*\|_2^2}{2\gamma_t} \\
\leqslant & \sum_{u=1}^{t}\frac{B^2\gamma_u}{2} + \sum_{u=1}^{t-1}4D^2\left(\frac{1}{2\gamma_{u+1}} - \frac{1}{2\gamma_u}\right) + \frac{4D^2}{2\gamma_1} \\
= & \sum_{u=1}^{t}\frac{B^2\gamma_u}{2} + \frac{4D^2}{2\gamma_t} \leqslant 3DB\sqrt{t} \text{ with } \gamma_t = \frac{2D}{B\sqrt{t}}
\end{aligned}$$

- Using convexity: $g\left(\frac{1}{t}\sum_{k=0}^{t-1}\theta_k\right) - g(\theta_*) \leqslant \frac{3DB}{\sqrt{t}}$

Subgradient descent for machine learning

- **Assumptions** (f is the expected risk, $\hat{f}$ the empirical risk)
  - "Linear" predictors: $\theta(x) = \theta^\top \Phi(x)$, with $\|\Phi(x)\|_2 \leqslant R$ a.s.
  - $\hat{f}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \Phi(x_i)^\top \theta)$
  - G-Lipschitz loss: f and $\hat{f}$ are GR-Lipschitz on $\Theta = \{\|\theta\|_2 \leqslant D\}$

- **Statistics**: with probability greater than $1 - \delta$

$$\sup_{\theta \in \Theta} |\hat{f}(\theta) - f(\theta)| \leqslant \frac{GRD}{\sqrt{n}} \left[2 + \sqrt{2 \log \frac{2}{\delta}} \right]$$

- **Optimization**: after t iterations of subgradient method

$$\hat{f}(\hat{\theta}) - \min_{\eta \in \Theta} \hat{f}(\eta) \leqslant \frac{GRD}{\sqrt{t}}$$

- $t = n$ iterations, with total running-time complexity of $O(n^2 d)$

Subgradient descent - strong convexity

- **Assumptions**
  - g convex and B-Lipschitz-continuous on $\{\|\theta\|_2 \leqslant D\}$
  - g μ-strongly convex
- **Algorithm**: $\theta_t = \Pi_D\left(\theta_{t-1} - \frac{2}{\mu(t+1)} g'(\theta_{t-1})\right)$
- **Bound**:
$$g\left(\frac{2}{t(t+1)}\sum_{k=1}^{t} k\theta_{k-1}\right) - g(\theta_*) \leqslant \frac{2B^2}{\mu(t+1)}$$
- Three-line proof
- Best possible convergence rate after $O(d)$ iterations (Bubeck, 2015)

Subgradient method - strong convexity - proof - I

- Iteration: $\theta_t = \Pi_D(\theta_{t-1} - \gamma_t g'(\theta_{t-1}))$ with $\gamma_t = \frac{2}{\mu(t+1)}$

- Assumption: $\|g'(\theta)\|_2 \leqslant B$ and $\|\theta\|_2 \leqslant D$ and μ-strong convexity of f

$$\begin{aligned}
\|\theta_t - \theta_*\|_2^2 &\leqslant \|\theta_{t-1} - \theta_* - \gamma_t g'(\theta_{t-1})\|_2^2 \text{ by contractivity of projections} \\
&\leqslant \|\theta_{t-1} - \theta_*\|_2^2 + B^2\gamma_t^2 - 2\gamma_t(\theta_{t-1} - \theta_*)^\top g'(\theta_{t-1}) \text{ because } \|g'(\theta_{t-1})\|_2 \leqslant \\
&\leqslant \|\theta_{t-1} - \theta_*\|_2^2 + B^2\gamma_t^2 - 2\gamma_t\big[g(\theta_{t-1}) - g(\theta_*) + \frac{\mu}{2}\|\theta_{t-1} - \theta_*\|_2^2\big]
\end{aligned}$$

(property of subgradients and strong convexity)

- leading to

$$\begin{aligned}
g(\theta_{t-1}) - g(\theta_*) &\leqslant \frac{B^2\gamma_t}{2} + \frac{1}{2}\big[\frac{1}{\gamma_t} - \mu\big]\|\theta_{t-1} - \theta_*\|_2^2 - \frac{1}{2\gamma_t}\|\theta_t - \theta_*\|_2^2 \\
&\leqslant \frac{B^2}{\mu(t+1)} + \frac{\mu}{2}\big[\frac{t-1}{2}\big]\|\theta_{t-1} - \theta_*\|_2^2 - \frac{\mu(t+1)}{4}\|\theta_t - \theta_*\|_2^2
\end{aligned}$$

Subgradient method - strong convexity - proof - II

- From $g(\theta_{t-1}) - g(\theta_*) \leqslant \frac{B^2}{\mu(t+1)} + \frac{\mu}{2}\big[\frac{t-1}{2}\big]\|\theta_{t-1} - \theta_*\|_2^2 - \frac{\mu(t+1)}{4}\|\theta_t - \theta_*\|_2^2$

$$\sum_{u=1}^{t} u\big[g(\theta_{u-1}) - g(\theta_*)\big] \leqslant \sum_{t=1}^{u} \frac{B^2 u}{\mu(u+1)} + \frac{1}{4}\sum_{u=1}^{t}\big[u(u-1)\|\theta_{u-1} - \theta_*\|_2^2 - u(u+1)\|\theta_u - \theta_*\|_2^2\big]$$

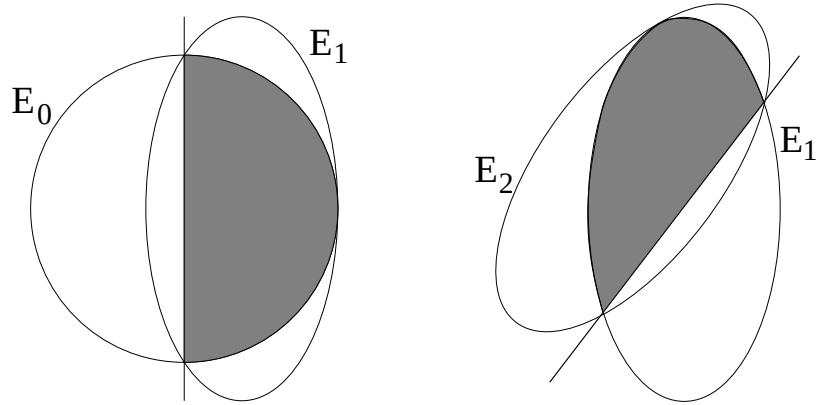
$$\leqslant \frac{B^2 t}{\mu} + \frac{1}{4}\big[0 - t(t+1)\|\theta_t - \theta_*\|_2^2\big] \leqslant \frac{B^2 t}{\mu}$$

- Using convexity: $g\left(\frac{2}{t(t+1)}\sum_{u=1}^{t} u\theta_{u-1}\right) - g(\theta_*) \leqslant \frac{2B^2}{t+1}$

- NB: with step-size $\gamma_n = 1/(n\mu)$, extra logarithmic factor

Ellipsoid method

- Minimizing convex function $g : \mathbb{R}^d \to \mathbb{R}$
  - Builds a sequence of ellipsoids that contains the global minima.



- Represent $E_t = \{\theta \in \mathbb{R}^d, (\theta - \theta_t)^\top P_t^{-1}(\theta - \theta_t) \leqslant 1\}$
- Fact 1: $\theta_{t+1} = \theta_t - \frac{1}{d+1} P_t h_t$ and $P_{t+1} = \frac{d^2}{d^2-1}(P_t - \frac{2}{d+1} P_t h_t h_t^\top P_t)$ with $h_t = \frac{1}{\sqrt{g'(\theta_t)^\top P_t g'(x_t)}} g'(\theta_t)$
- Fact 2: $\mathrm{vol}(\mathcal{E}_t) \approx \mathrm{vol}(\mathcal{E}_{t-1}) e^{-1/2d} \Rightarrow$ CV rate in $O(e^{-t/d^2})$

Summary: minimizing convex functions

- **Gradient descent**: $\theta_t = \theta_{t-1} - \gamma_t \, g'(\theta_{t-1})$
  - $O(1/\sqrt{t})$ convergence rate for non-smooth convex functions
  - $O(1/t)$ convergence rate for smooth convex functions
  - $O(e^{-\rho t})$ convergence rate for strongly smooth convex functions

- **Newton method**: $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1} g'(\theta_{t-1})$
  - $O\left(e^{-\rho 2^t}\right)$ convergence rate

Summary: minimizing convex functions

- **Gradient descent**: $\theta_t = \theta_{t-1} - \gamma_t \, g'(\theta_{t-1})$
  - $O(1/\sqrt{t})$ convergence rate for non-smooth convex functions
  - $O(1/t)$ convergence rate for smooth convex functions
  - $O(e^{-\rho t})$ convergence rate for strongly smooth convex functions

- **Newton method**: $\theta_t = \theta_{t-1} - g''(\theta_{t-1})^{-1} g'(\theta_{t-1})$
  - $O\left(e^{-\rho 2^t}\right)$ convergence rate

- **Key insights from Bottou and Bousquet (2008)**

1. In machine learning, no need to optimize below statistical error
2. In machine learning, cost functions are averages
3. Testing errors are more important than training errors

$\Rightarrow$ **Stochastic approximation**

Summary of rates of convergence

- Problem parameters
  - D diameter of the domain
  - B Lipschitz-constant
  - L smoothness constant
  - μ strong convexity constant

	convex	strongly convex
nonsmooth	deterministic: $BD/\sqrt{t}$	deterministic: $B^2/(t\mu)$
smooth	deterministic: LD^2/t^2	deterministic: $\exp(-t\sqrt{\mu/L})$
quadratic	deterministic: LD^2/t^2	deterministic: $\exp(-t\sqrt{\mu/L})$