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

Stochastic approximation

- **Goal**: Minimizing a function f defined on $\mathbb{R}^d$
  - given only unbiased estimates $f'_n(\theta_n)$ of its gradients $f'(\theta_n)$ at certain points $\theta_n \in \mathbb{R}^d$

Stochastic approximation

- **Goal**: Minimizing a function f defined on $\mathbb{R}^d$
  - given only unbiased estimates $f_n'(\theta_n)$ of its gradients $f'(\theta_n)$ at certain points $\theta_n \in \mathbb{R}^d$

- **Machine learning - statistics**
  - **loss for a single pair of observations**: $\boxed{f_n(\theta) = \ell(y_n, \theta^\top \Phi(x_n))}$
  - $f(\theta) = \mathbb{E} f_n(\theta) = \mathbb{E}\, \ell(y_n, \theta^\top \Phi(x_n)) =$ **generalization error**
  - Expected gradient: $f'(\theta) = \mathbb{E} f_n'(\theta) = \mathbb{E}\left\{\ell'(y_n, \theta^\top \Phi(x_n))\, \Phi(x_n)\right\}$
  - Non-asymptotic results

- **Number of iterations = number of observations**

Stochastic approximation

- **Goal**: Minimizing a function f defined on $\mathbb{R}^d$
  - given only unbiased estimates $f'_n(\theta_n)$ of its gradients $f'(\theta_n)$ at certain points $\theta_n \in \mathbb{R}^d$

- **Stochastic approximation**
  - (much) broader applicability beyond convex optimization

$$\theta_n = \theta_{n-1} - \gamma_n h_n(\theta_{n-1}) \text{ with } \mathbb{E}\big[h_n(\theta_{n-1})|\theta_{n-1}\big] = h(\theta_{n-1})$$

  - Beyond convex problems, i.i.d assumption, finite dimension, etc.
  - Typically asymptotic results (see next lecture)
  - See, e.g., Kushner and Yin (2003); Benveniste et al. (2012)

Relationship to online learning

- **Stochastic approximation**
  - Minimize $f(\theta) = \mathbb{E}_z \ell(\theta, z) =$ **generalization error** of θ
  - Using the gradients of single i.i.d. observations

Relationship to online learning

- **Stochastic approximation**
  - Minimize $f(\theta) = \mathbb{E}_z \ell(\theta, z) =$ **generalization error** of θ
  - Using the gradients of single i.i.d. observations

- **Batch learning**
  - Finite set of observations: $z_1, \ldots, z_n$
  - Empirical risk: $\hat{f}(\theta) = \frac{1}{n} \sum_{k=1}^{n} \ell(\theta, z_i)$
  - Estimator $\hat{\theta} =$ Minimizer of $\hat{f}(\theta)$ over a certain class Θ
  - Generalization bound using uniform concentration results

Relationship to online learning

- **Stochastic approximation**
  - Minimize $f(\theta) = \mathbb{E}_z \ell(\theta, z) =$ **generalization error** of θ
  - Using the gradients of single i.i.d. observations
- **Batch learning**
  - Finite set of observations: $z_1, \ldots, z_n$
  - Empirical risk: $\hat{f}(\theta) = \frac{1}{n} \sum_{k=1}^{n} \ell(\theta, z_i)$
  - Estimator $\hat{\theta} =$ Minimizer of $\hat{f}(\theta)$ over a certain class Θ
  - Generalization bound using uniform concentration results
- **Online learning**
  - Update $\hat{\theta}_n$ after each new (potentially adversarial) observation z_n
  - Cumulative loss: $\frac{1}{n} \sum_{k=1}^{n} \ell(\hat{\theta}_{k-1}, z_k)$
  - Online to batch through averaging (Cesa-Bianchi et al., 2004)

Convex stochastic approximation

- **Key properties of f and/or f_n**
  - Smoothness: f B-Lipschitz continuous, f' L-Lipschitz continuous
  - Strong convexity: f μ-strongly convex

Convex stochastic approximation

- **Key properties of f and/or f_n**
  - Smoothness: f B-Lipschitz continuous, f' L-Lipschitz continuous
  - Strong convexity: f μ-strongly convex

- **Key algorithm:** Stochastic gradient descent (a.k.a. Robbins-Monro)

$$\boxed{\theta_n = \theta_{n-1} - \gamma_n f_n'(\theta_{n-1})}$$

  - Polyak-Ruppert averaging: $\bar{\theta}_n = \frac{1}{n} \sum_{k=0}^{n-1} \theta_k$
  - Which learning rate sequence γ_n? Classical setting: $\boxed{\gamma_n = Cn^{-\alpha}}$

Convex stochastic approximation

- **Key properties of f and/or f_n**
  - Smoothness: f B-Lipschitz continuous, f' L-Lipschitz continuous
  - Strong convexity: f μ-strongly convex
- **Key algorithm:** Stochastic gradient descent (a.k.a. Robbins-Monro)

$$\boxed{\theta_n = \theta_{n-1} - \gamma_n f_n'(\theta_{n-1})}$$

  - Polyak-Ruppert averaging: $\bar{\theta}_n = \frac{1}{n} \sum_{k=0}^{n-1} \theta_k$
  - Which learning rate sequence γ_n? Classical setting: $\boxed{\gamma_n = Cn^{-\alpha}}$
- **Desirable practical behavior**
  - Applicable (at least) to classical supervised learning problems
  - Robustness to (potentially unknown) constants (L,B,μ)
  - Adaptivity to difficulty of the problem (e.g., strong convexity)

Stochastic subgradient "descent"/method

- **Assumptions**
  - f_n convex and B-Lipschitz-continuous on $\{\|\theta\|_2 \leqslant D\}$
  - (f_n) i.i.d. functions such that $\mathbb{E}f_n = f$
  - θ_* global optimum of f on $\mathcal{C} = \{\|\theta\|_2 \leqslant D\}$

- **Algorithm**: $\theta_n = \Pi_D\left(\theta_{n-1} - \dfrac{2D}{B\sqrt{n}} f_n'(\theta_{n-1})\right)$

Stochastic subgradient "descent"/method

- **Assumptions**
  - f_n convex and B-Lipschitz-continuous on $\{\|\theta\|_2 \leqslant D\}$
  - (f_n) i.i.d. functions such that $\mathbb{E}f_n = f$
  - θ_* global optimum of f on $\mathcal{C} = \{\|\theta\|_2 \leqslant D\}$
- **Algorithm**: $\theta_n = \Pi_D\left(\theta_{n-1} - \dfrac{2D}{B\sqrt{n}} f_n'(\theta_{n-1})\right)$
- **Bound**:

$$\mathbb{E}f\left(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\right) - f(\theta_*) \leqslant \frac{2DB}{\sqrt{n}}$$

- "Same" three-line proof as in the deterministic case
- Minimax rate (Nemirovsky and Yudin, 1983; Agarwal et al., 2012)
- Running-time complexity: $O(dn)$ after n iterations

Stochastic subgradient method - proof - I

- Iteration: $\theta_n = \Pi_D(\theta_{n-1} - \gamma_n f_n'(\theta_{n-1}))$ with $\gamma_n = \frac{2D}{B\sqrt{n}}$
- $\mathcal{F}_n$: information up to time n
- $\|f_n'(\theta)\|_2 \leqslant B$ and $\|\theta\|_2 \leqslant D$, unbiased gradients/functions $\mathbb{E}(f_n|\mathcal{F}_{n-1}) = f$

$$\begin{aligned}\|\theta_n - \theta_*\|_2^2 &\leqslant \|\theta_{n-1} - \theta_* - \gamma_n f_n'(\theta_{n-1})\|_2^2 \text{ by contractivity of projections} \\ &\leqslant \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta_*)^\top f_n'(\theta_{n-1}) \text{ because } \|f_n'(\theta_{n-1})\|_2\end{aligned}$$

$$\begin{aligned}\mathbb{E}\big[\|\theta_n - \theta_*\|_2^2|\mathcal{F}_{n-1}\big] &\leqslant \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta_*)^\top f'(\theta_{n-1}) \\ &\leqslant \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n\big[f(\theta_{n-1}) - f(\theta_*)\big] \text{ (subgradient prope} \\ \mathbb{E}\|\theta_n - \theta_*\|_2^2 &\leqslant \mathbb{E}\|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n\big[\mathbb{E}f(\theta_{n-1}) - f(\theta_*)\big]\end{aligned}$$

- leading to $\mathbb{E}f(\theta_{n-1}) - f(\theta_*) \leqslant \dfrac{B^2\gamma_n}{2} + \dfrac{1}{2\gamma_n}\big[\mathbb{E}\|\theta_{n-1} - \theta_*\|_2^2 - \mathbb{E}\|\theta_n - \theta_*\|_2^2\big]$

Stochastic subgradient method - proof - II

- Starting from $\mathbb{E}f(\theta_{n-1}) - f(\theta_*) \leqslant \frac{B^2\gamma_n}{2} + \frac{1}{2\gamma_n}\left[\mathbb{E}\|\theta_{n-1} - \theta_*\|_2^2 - \mathbb{E}\|\theta_n - \theta_*\|_2^2\right]$

$$\begin{aligned}\sum_{u=1}^{n}\left[\mathbb{E}f(\theta_{u-1}) - f(\theta_*)\right] \leqslant & \sum_{u=1}^{n}\frac{B^2\gamma_u}{2} + \sum_{u=1}^{n}\frac{1}{2\gamma_u}\left[\mathbb{E}\|\theta_{u-1} - \theta_*\|_2^2 - \mathbb{E}\|\theta_u - \theta_*\|_2^2\right] \\ \leqslant & \sum_{u=1}^{n}\frac{B^2\gamma_u}{2} + \frac{4D^2}{2\gamma_n} \leqslant 2DB\sqrt{n} \text{ with } \gamma_n = \frac{2D}{B\sqrt{n}}\end{aligned}$$

- Using convexity: $\mathbb{E}f\left(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\right) - f(\theta_*) \leqslant \frac{2DB}{\sqrt{n}}$

Stochastic subgradient method
Extension to online learning

- Assume different and arbitrary functions $f_n : \mathbb{R}^d \to \mathbb{R}$
  - Observations of $f_n'(\theta_{n-1}) + \varepsilon_n$
  - with $\mathbb{E}(\varepsilon_n | \mathcal{F}_{n-1}) = 0$ and $\|f_n'(\theta_{n-1}) + \varepsilon_n\| \leqslant B$ almost surely

- **Performance criterion**: (normalized) regret

$$\frac{1}{n} \sum_{i=1}^{n} f_i(\theta_{i-1}) - \inf_{\|\theta\|_2 \leqslant D} \frac{1}{n} \sum_{i=1}^{n} f_i(\theta)$$

  - Warning: often not normalized
  - May not be non-negative (typically is)

Stochastic subgradient method - online learning - I

- Iteration: $\theta_n = \Pi_D(\theta_{n-1} - \gamma_n(f_n'(\theta_{n-1}) + \varepsilon_n))$ with $\gamma_n = \frac{2D}{B\sqrt{n}}$
- $\mathcal{F}_n$: information up to time n - θ an arbitrary point such that $\|\theta\| \leqslant D$
- $\|f_n'(\theta_{n-1}) + \varepsilon_n\|_2 \leqslant B$ and $\|\theta\|_2 \leqslant D$, unbiased gradients $\mathbb{E}(\varepsilon_n|\mathcal{F}_{n-1}) = 0$

$$\|\theta_n - \theta\|_2^2 \leqslant \|\theta_{n-1} - \theta - \gamma_n(f_n'(\theta_{n-1}) + \varepsilon_n)\|_2^2 \text{ by contractivity of projections}$$
$$\leqslant \|\theta_{n-1} - \theta\|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta)^\top(f_n'(\theta_{n-1}) + \varepsilon_n) \text{ because } \|f_n'(\theta_{n-1}) +$$

$$\mathbb{E}\big[\|\theta_n - \theta\|_2^2|\mathcal{F}_{n-1}\big] \leqslant \|\theta_{n-1} - \theta\|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta)^\top f_n'(\theta_{n-1})$$
$$\leqslant \|\theta_{n-1} - \theta\|_2^2 + B^2\gamma_n^2 - 2\gamma_n\big[f_n(\theta_{n-1}) - f_n(\theta)\big] \text{ (subgradient proper}$$
$$\mathbb{E}\|\theta_n - \theta\|_2^2 \leqslant \mathbb{E}\|\theta_{n-1} - \theta\|_2^2 + B^2\gamma_n^2 - 2\gamma_n\big[\mathbb{E}f_n(\theta_{n-1}) - f_n(\theta)\big]$$

- leading to $\mathbb{E}f_n(\theta_{n-1}) - f_n(\theta) \leqslant \dfrac{B^2\gamma_n}{2} + \dfrac{1}{2\gamma_n}\big[\mathbb{E}\|\theta_{n-1} - \theta\|_2^2 - \mathbb{E}\|\theta_n - \theta\|_2^2\big]$

Stochastic subgradient method - online learning - II

- Starting from $\mathbb{E}f_n(\theta_{n-1}) - f_n(\theta) \leqslant \frac{B^2\gamma_n}{2} + \frac{1}{2\gamma_n}\big[\mathbb{E}\|\theta_{n-1} - \theta\|_2^2 - \mathbb{E}\|\theta_n - \theta\|_2^2\big]$

$$\begin{aligned}\sum_{u=1}^{n}\big[\mathbb{E}f_u(\theta_{u-1}) - f_u(\theta)\big] \leqslant & \sum_{u=1}^{n}\frac{B^2\gamma_u}{2} + \sum_{u=1}^{n}\frac{1}{2\gamma_u}\big[\mathbb{E}\|\theta_{u-1} - \theta\|_2^2 - \mathbb{E}\|\theta_u - \theta\|_2^2\big] \\ & \leqslant \sum_{u=1}^{n}\frac{B^2\gamma_u}{2} + \frac{4D^2}{2\gamma_n} \leqslant 2DB\sqrt{n} \text{ with } \gamma_n = \frac{2D}{B\sqrt{n}}\end{aligned}$$

- For any θ such that $\|\theta\| \leqslant D$: $\displaystyle\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}f_k(\theta_{k-1}) - \frac{1}{n}\sum_{k=1}^{n}f_k(\theta) \leqslant \frac{2DB}{\sqrt{n}}$
- Online to batch conversion: assuming convexity

Stochastic subgradient descent - strong convexity - I

- **Assumptions**
  - f_n convex and B-Lipschitz-continuous
  - (f_n) i.i.d. functions such that $\mathbb{E}f_n = f$
  - f μ-strongly convex on $\{\|\theta\|_2 \leqslant D\}$
  - θ_* global optimum of f over $\{\|\theta\|_2 \leqslant D\}$

- **Algorithm**: $\theta_n = \Pi_D\left(\theta_{n-1} - \dfrac{2}{\mu(n+1)} f_n'(\theta_{n-1})\right)$

- **Bound**:

$$\mathbb{E}f\left(\frac{2}{n(n+1)}\sum_{k=1}^{n} k\theta_{k-1}\right) - f(\theta_*) \leqslant \frac{2B^2}{\mu(n+1)}$$

- "Same" proof than deterministic case (Lacoste-Julien et al., 2012)

- Minimax rate (Nemirovsky and Yudin, 1983; Agarwal et al., 2012)

Stochastic subgradient - strong convexity - proof - I

- Iteration: $\theta_n = \Pi_D(\theta_{n-1} - \gamma_n f_n'(\theta_{t-1}))$ with $\gamma_n = \frac{2}{\mu(n+1)}$

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

$$
\begin{aligned}
\|\theta_n - \theta_*\|_2^2 &\leqslant \|\theta_{n-1} - \theta_* - \gamma_n f_n'(\theta_{t-1})\|_2^2 \text{ by contractivity of projections} \\
&\leqslant \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta_*)^\top f_n'(\theta_{t-1}) \text{ because } \|f_n'(\theta_{t-1})\|_2 \\
\mathbb{E}(\cdot|\mathcal{F}_{n-1}) &\leqslant \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n\big[f(\theta_{n-1}) - f(\theta_*) + \frac{\mu}{2}\|\theta_{n-1} - \theta_*\|_2^2\big]
\end{aligned}
$$

(property of subgradients and strong convexity)

- leading to

$$
\begin{aligned}
\mathbb{E}f(\theta_{n-1}) - f(\theta_*) &\leqslant \frac{B^2\gamma_n}{2} + \frac{1}{2}\big[\frac{1}{\gamma_n} - \mu\big]\|\theta_{n-1} - \theta_*\|_2^2 - \frac{1}{2\gamma_n}\|\theta_n - \theta_*\|_2^2 \\
&\leqslant \frac{B^2}{\mu(n+1)} + \frac{\mu}{2}\big[\frac{n-1}{2}\big]\|\theta_{n-1} - \theta_*\|_2^2 - \frac{\mu(n+1)}{4}\|\theta_n - \theta_*\|_2^2
\end{aligned}
$$

Stochastic subgradient - strong convexity - proof - II

- From $\mathbb{E}f(\theta_{n-1}) - f(\theta_*) \leqslant \frac{B^2}{\mu(n+1)} + \frac{\mu}{2}\big[\frac{n-1}{2}\big]\mathbb{E}\|\theta_{n-1} - \theta_*\|_2^2 - \frac{\mu(n+1)}{4}\mathbb{E}\|\theta_n - \theta_*\|_2^2$

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

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

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

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

Stochastic subgradient descent - strong convexity - II

- **Assumptions**
  - f_n convex and B-Lipschitz-continuous
  - (f_n) i.i.d. functions such that $\mathbb{E}f_n = f$
  - θ_* global optimum of $g = f + \frac{\mu}{2}\|\cdot\|_2^2$
  - No compactness assumption - no projections

- **Algorithm**:

$$\theta_n = \theta_{n-1} - \frac{2}{\mu(n+1)} g_n'(\theta_{n-1}) = \theta_{n-1} - \frac{2}{\mu(n+1)}\left[f_n'(\theta_{n-1}) + \mu\theta_{n-1}\right]$$

- **Bound**: $\mathbb{E}g\left(\frac{2}{n(n+1)}\sum_{k=1}^{n} k\theta_{k-1}\right) - g(\theta_*) \leqslant \frac{2B^2}{\mu(n+1)}$

- Minimax convergence rate

Strong convexity - proof with $\log n$ factor - I

- Iteration: $\theta_n = \Pi_D(\theta_{n-1} - \gamma_n f_n'(\theta_{t-1}))$ with $\gamma_n = \frac{1}{\mu n}$

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

$$\begin{aligned}
\|\theta_n - \theta_*\|_2^2 &\leqslant \|\theta_{n-1} - \theta_* - \gamma_n f_n'(\theta_{t-1})\|_2^2 \text{ by contractivity of projections} \\
&\leqslant \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta_*)^\top f_n'(\theta_{t-1}) \text{ because } \|f_n'(\theta_{t-1})\|_2 \\
\mathbb{E}(\cdot|\mathcal{F}_{n-1}) &\leqslant \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n\big[f(\theta_{n-1}) - f(\theta_*) + \frac{\mu}{2}\|\theta_{n-1} - \theta_*\|_2^2\big]
\end{aligned}$$

(property of subgradients and strong convexity)

- leading to

$$\begin{aligned}
\mathbb{E}f(\theta_{n-1}) - f(\theta_*) &\leqslant \frac{B^2\gamma_n}{2} + \frac{1}{2}\big[\frac{1}{\gamma_n} - \mu\big]\|\theta_{n-1} - \theta_*\|_2^2 - \frac{1}{2\gamma_n}\|\theta_n - \theta_*\|_2^2 \\
&\leqslant \frac{B^2}{2\mu n} + \frac{\mu}{2}\big[n-1\big]\|\theta_{n-1} - \theta_*\|_2^2 - \frac{n\mu}{2}\|\theta_n - \theta_*\|_2^2
\end{aligned}$$

Strong convexity - proof with $\log n$ factor - II

- From $\mathbb{E}f(\theta_{n-1}) - f(\theta_*) \leqslant \frac{B^2}{2\mu n} + \frac{\mu}{2}\big[n-1\big]\|\theta_{n-1} - \theta_*\|_2^2 - \frac{n\mu}{2}\|\theta_n - \theta_*\|_2^2$

$$\sum_{u=1}^{n}\big[\mathbb{E}f(\theta_{u-1}) - f(\theta_*)\big] \leqslant \sum_{u=1}^{n}\frac{B^2}{2\mu u} + \frac{1}{2}\sum_{u=1}^{n}\big[(u-1)\mathbb{E}\|\theta_{u-1} - \theta_*\|_2^2 - u\mathbb{E}\|\theta_u - \theta_*\|_2^2\big]$$
$$\leqslant \frac{B^2\log n}{2\mu} + \frac{1}{2}\big[0 - n\mathbb{E}\|\theta_n - \theta_*\|_2^2\big] \leqslant \frac{B^2\log n}{2\mu}$$

- Using convexity: $\mathbb{E}f\left(\frac{1}{n}\sum_{u=1}^{n}\theta_{u-1}\right) - f(\theta_*) \leqslant \frac{B^2}{2\mu}\frac{\log n}{n}$
- Why could this be useful?

Stochastic subgradient descent - strong convexity
Online learning

- Need $\log n$ term for uniform averaging. For all θ:

$$\frac{1}{n}\sum_{i=1}^{n} f_i(\theta_{i-1}) - \frac{1}{n}\sum_{i=1}^{n} f_i(\theta) \leqslant \frac{B^2}{2\mu}\frac{\log n}{n}$$

- Optimal. See Hazan and Kale (2014).

Beyond convergence in expectation

- **Typical result**: $\mathbb{E} f\left(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\right) - f(\theta_*) \leqslant \frac{2DB}{\sqrt{n}}$
  - Obtained with simple conditioning arguments
- **High-probability bounds**
  - Markov inequality: $\mathbb{P}\left(f\left(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\right) - f(\theta_*) \geqslant \varepsilon\right) \leqslant \frac{2DB}{\sqrt{n}\varepsilon}$

Beyond convergence in expectation

- **Typical result**: $\mathbb{E} f\left(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\right) - f(\theta_*) \leqslant \frac{2DB}{\sqrt{n}}$
  - Obtained with simple conditioning arguments
- **High-probability bounds**
  - Markov inequality: $\mathbb{P}\left(f\left(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\right) - f(\theta_*) \geqslant \varepsilon\right) \leqslant \frac{2DB}{\sqrt{n}\varepsilon}$
  - Deviation inequality (Nemirovski et al., 2009; Nesterov and Vial, 2008)

$$\mathbb{P}\left(f\left(\frac{1}{n}\sum_{k=0}^{n-1}\theta_k\right) - f(\theta_*) \geqslant \frac{2DB}{\sqrt{n}}(2+4t)\right) \leqslant 2\exp(-t^2)$$

- See also Bach (2013) for logistic regression

Stochastic subgradient method - high probability - I

- Iteration: $\theta_n = \Pi_D(\theta_{n-1} - \gamma_n f_n'(\theta_{n-1}))$ with $\gamma_n = \frac{2D}{B\sqrt{n}}$
- $\mathcal{F}_n$: information up to time n
- $\|f_n'(\theta)\|_2 \leqslant B$ and $\|\theta\|_2 \leqslant D$, unbiased gradients/functions $\mathbb{E}(f_n|\mathcal{F}_{n-1}) = f$

$$\begin{aligned}\|\theta_n - \theta_*\|_2^2 &\leqslant \|\theta_{n-1} - \theta_* - \gamma_n f_n'(\theta_{n-1})\|_2^2 \text{ by contractivity of projections} \\ &\leqslant \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta_*)^\top f_n'(\theta_{n-1}) \text{ because } \|f_n'(\theta_{n-1})\|_2\end{aligned}$$

$$\begin{aligned}\mathbb{E}\big[\|\theta_n - \theta_*\|_2^2|\mathcal{F}_{n-1}\big] &\leqslant \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n(\theta_{n-1} - \theta_*)^\top f'(\theta_{n-1}) \\ &\leqslant \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n\big[f(\theta_{n-1}) - f(\theta_*)\big] \text{ (subgradient prope}\end{aligned}$$

- Without expectations and with $Z_n = -2\gamma_n(\theta_{n-1} - \theta_*)^\top[f_n'(\theta_{n-1}) - f'(\theta_{n-1})]$

$$\|\theta_n - \theta_*\|_2^2 \leqslant \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n\big[f(\theta_{n-1}) - f(\theta_*)\big] + Z_n$$

Stochastic subgradient method - high probability - II

- Without expectations and with $Z_n = -2\gamma_n(\theta_{n-1} - \theta_*)^\top [f'_n(\theta_{n-1}) - f'(\theta_{n-1})]$

$$\begin{aligned}
\|\theta_n - \theta_*\|_2^2 &\leqslant \|\theta_{n-1} - \theta_*\|_2^2 + B^2\gamma_n^2 - 2\gamma_n\big[f(\theta_{n-1}) - f(\theta_*)\big] + Z_n \\
f(\theta_{n-1}) - f(\theta_*) &\leqslant \frac{1}{2\gamma_n}\big[\|\theta_{n-1} - \theta_*\|_2^2 - \|\theta_n - \theta_*\|_2^2\big] + \frac{B^2\gamma_n}{2} + \frac{Z_n}{2\gamma_n}
\end{aligned}$$

$$\begin{aligned}
\sum_{u=1}^n \big[f(\theta_{u-1}) - f(\theta_*)\big] &\leqslant \sum_{u=1}^n \frac{B^2\gamma_u}{2} + \sum_{u=1}^n \frac{1}{2\gamma_u}\big[\|\theta_{u-1} - \theta_*\|_2^2 - \|\theta_u - \theta_*\|_2^2\big] + \sum_{u=1}^n \frac{Z_u}{2\gamma_u} \\
&\leqslant \sum_{u=1}^n \frac{B^2\gamma_u}{2} + \frac{4D^2}{2\gamma_n} + \sum_{u=1}^n \frac{Z_u}{2\gamma_u} \leqslant \frac{2DB}{\sqrt{n}} + \sum_{u=1}^n \frac{Z_u}{2\gamma_u} \text{ with } \gamma_n = \frac{2D}{B\sqrt{n}}
\end{aligned}$$

- Need to study $\displaystyle\sum_{u=1}^n \frac{Z_u}{2\gamma_u}$ with $\mathbb{E}(Z_n|\mathcal{F}_{n-1}) = 0$ and $|Z_n| \leqslant 8\gamma_n DB$

Stochastic subgradient method - high probability - III

- Need to study $\sum_{u=1}^{n} \frac{Z_u}{2\gamma_u}$ with $\mathbb{E}(\frac{Z_n}{2\gamma_n}|\mathcal{F}_{n-1}) = 0$ and $|Z_n| \leqslant 4DB$

- Azuma-Hoeffding inequality for bounded martingale increments:

$$\mathbb{P}\Big(\sum_{u=1}^{n} \frac{Z_u}{2\gamma_u} \geqslant t\sqrt{n} \cdot 4DB\Big) \leqslant \exp\Big(-\frac{t^2}{2}\Big)$$

- Moments with Burkholder-Rosenthal-Pinelis inequality (Pinelis, 1994)

Beyond stochastic gradient method

- **Adding a proximal step**
  - Goal: $\min_{\theta \in \mathbb{R}^d} f(\theta) + \Omega(\theta) = \mathbb{E} f_n(\theta) + \Omega(\theta)$
  - Replace recursion $\theta_n = \theta_{n-1} - \gamma_n f_n'(\theta_n)$ by

$$\theta_n = \min_{\theta \in \mathbb{R}^d} \left\| \theta - \theta_{n-1} + \gamma_n f_n'(\theta_n) \right\|_2^2 + C\Omega(\theta)$$

  - Xiao (2010); Hu et al. (2009)
  - May be accelerated (Ghadimi and Lan, 2013)

- **Related frameworks**
  - Regularized dual averaging (Nesterov, 2009; Xiao, 2010)
  - Mirror descent (Nemirovski et al., 2009; Lan et al., 2012)

Mirror descent

- Projected (stochastic) gradient descent adapted to Euclidean geometry
  - bound: $\frac{\max_{\theta,\theta' \in \Theta} \|\theta - \theta'\|_2 \cdot \max_{\theta \in \Theta} \|f'(\theta)\|_2}{\sqrt{n}}$
- What about other norms?
  - Example: natural bound on $\max_{\theta \in \Theta} \|f'(\theta)\|_\infty$ leads to $\sqrt{d}$ factor
  - Avoidable with mirror descent, which leads to factor $\sqrt{\log d}$
  - Nemirovski et al. (2009); Lan et al. (2012)

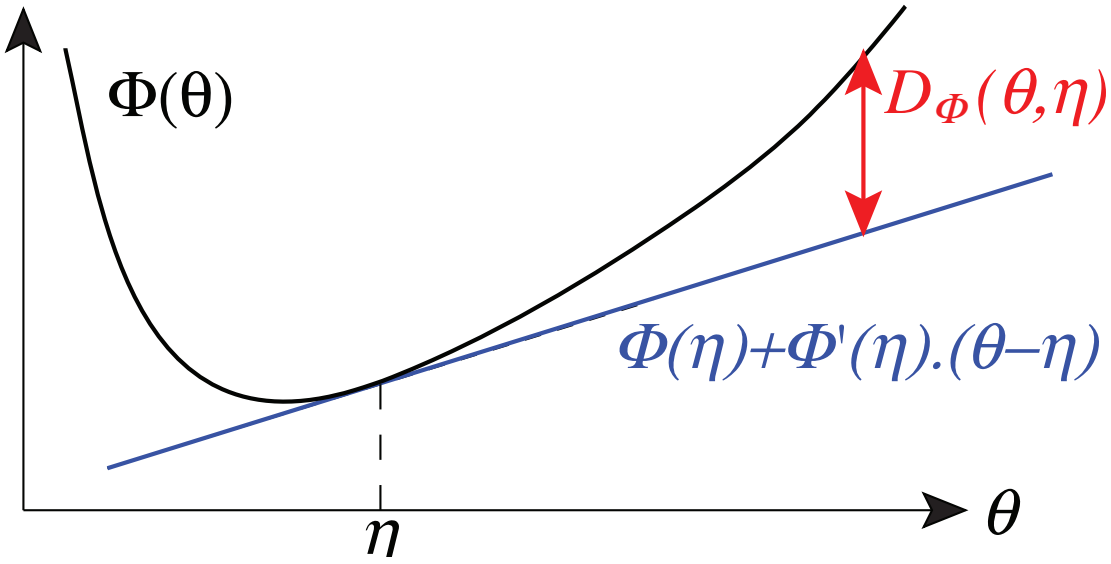
Mirror descent

- Projected (stochastic) gradient descent adapted to Euclidean geometry
  - bound: $\frac{\max_{\theta,\theta' \in \Theta} \|\theta - \theta'\|_2 \cdot \max_{\theta \in \Theta} \|f'(\theta)\|_2}{\sqrt{n}}$
- What about other norms?
  - Example: natural bound on $\max_{\theta \in \Theta} \|f'(\theta)\|_\infty$ leads to $\sqrt{d}$ factor
  - Avoidable with mirror descent, which leads to factor $\sqrt{\log d}$
  - Nemirovski et al. (2009); Lan et al. (2012)
- From Hilbert to Banach spaces
  - Gradient $f'(\theta)$ defined through $f(\theta + d\theta) - f(\theta) = \langle f'(\theta), d\theta \rangle$ for a certain dot-product
  - Generally, the differential is an element of the dual space

Mirror descent set-up

- Function f defined on domain $\mathcal{C}$
- Arbitrary norm $\|\cdot\|$ with dual norm $\|s\|_* = \sup_{\|\theta\| \leqslant 1} \theta^\top s$
- B-Lipschitz-continuous function w.r.t. $\|\cdot\|$: $\|f'(\theta)\|_* \leqslant B$
- Given a strictly-convex function Φ, define the Bregman divergence

$$D_\Phi(\theta, \eta) = \Phi(\theta) - \Phi(\eta) - \Phi'(\eta)^\top(\theta - \eta)$$



Mirror map

- Strongly-convex function $\Phi : \mathcal{C}_\Phi \to \mathbb{R}$ such that

(a) the gradient Φ' takes all possible values in $\mathbb{R}^d$, leading to a bijection from $\mathcal{C}_\Phi$ to $\mathbb{R}^d$

(b) the gradient Φ' diverges on the boundary of $\mathcal{C}_\Phi$

(c) $\mathcal{C}_\Phi$ contains the closure of the domain $\mathcal{C}$ of the optimization problem

- Bregman projection on $\mathcal{C}$ uniquely defined on $\mathcal{C}_\Phi$:

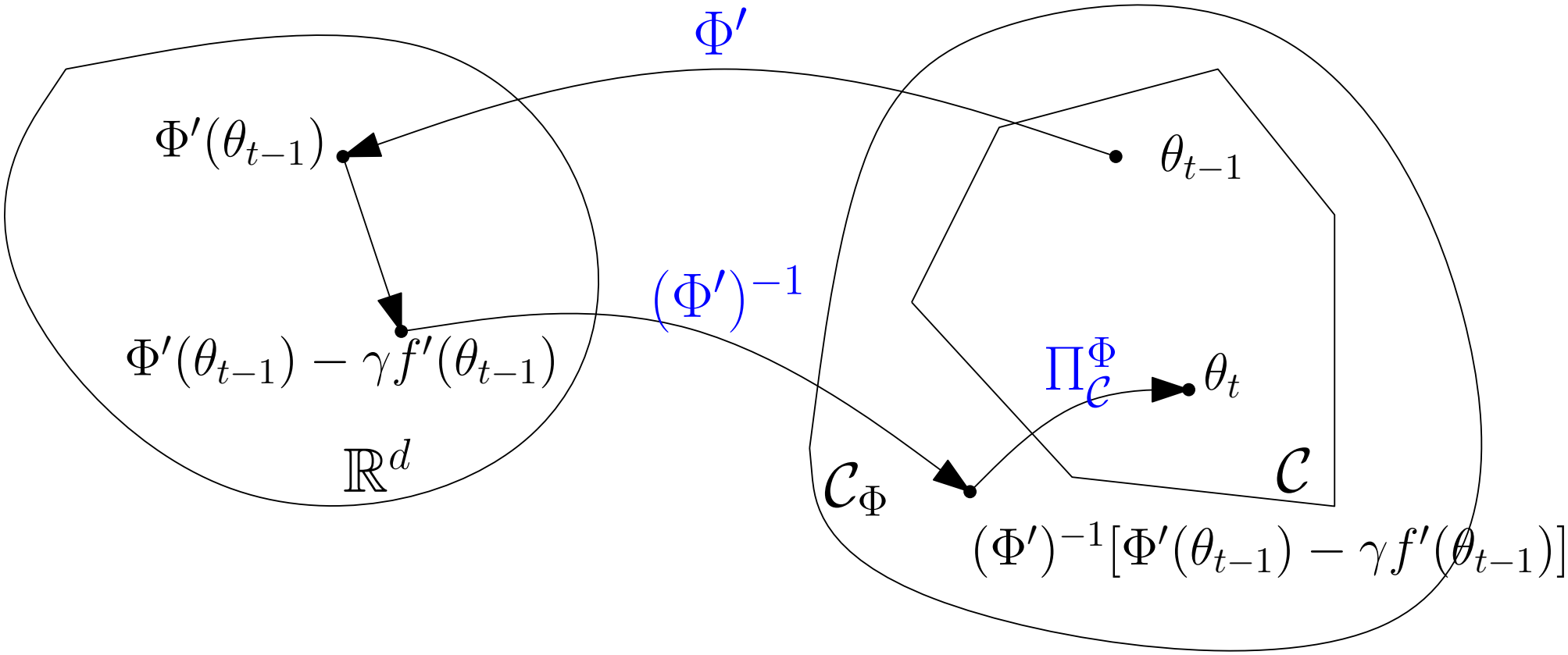
$$\begin{aligned}
\Pi_{\mathcal{C}}^{\Phi}(\theta) &= \arg\min_{\eta \in \mathcal{C}_\Phi \cap \mathcal{C}} D_\Phi(\eta, \theta) \\
&= \arg\min_{\eta \in \mathcal{C}_\Phi \cap \mathcal{C}} \Phi(\eta) - \Phi(\theta) - \Phi'(\theta)^\top(\eta - \theta) \\
&= \arg\min_{\eta \in \mathcal{C}_\Phi \cap \mathcal{C}} \Phi(\eta) - \Phi'(\theta)^\top \eta
\end{aligned}$$

- Example of squared Euclidean norm and entropy

Mirror descent

- **Iteration**:

$$\theta_t \;=\; \Pi_{\mathcal{C}}^{\Phi}\big(\Phi'^{-1}\big[\Phi'(\theta_{t-1}) - \gamma f'(\theta_{t-1})\big]\big)$$



Mirror descent

- **Iteration**:

$$\theta_t \;=\; \Pi_{\mathcal{C}}^{\Phi}\big(\Phi'^{-1}\big[\Phi'(\theta_{t-1}) - \gamma f'(\theta_{t-1})\big]\big)$$

- **Convergence**: assume (a) $D^2 = \sup_{\theta\in\mathcal{C}} \Phi(\theta) - \inf_{\theta\in\mathcal{C}} \Phi(\theta)$, (b) Φ is α-strongly convex with respect to $\|\cdot\|$ and (c) f is B-Lipschitz-continuous wr.t. $\|\cdot\|$. Then with $\gamma = \frac{D}{B}\sqrt{\frac{2\alpha}{t}}$:

$$f\left(\frac{1}{t}\sum_{u=1}^{t}\theta_u\right) - \inf_{\theta\in\mathcal{C}} f(\theta) \leqslant DB\sqrt{\frac{2}{\alpha t}}$$

  - See detailed proof in Bubeck (2015, p. 299)
  - "Same" as subgradient method + allows stochastic gradients

Mirror descent (proof)

- Define $\Phi'(\eta_t) = \Phi'(\theta_{t-1}) - \gamma f'(\theta_{t-1})$. We have

$$\begin{aligned} f(\theta_{t-1}) - f(\theta) &\leqslant f'(\theta_{t-1})^\top(\theta_{t-1} - \theta) = \frac{1}{\gamma}(\Phi'(\theta_{t-1}) - \Phi'(\eta_t))^\top(\theta_{t-1} - \theta) \\ &= \frac{1}{\gamma}\big[D_\Phi(\theta, \theta_{t-1}) + D_\Phi(\theta_{t-1}, \eta_t) - D_\Phi(\theta, \eta_t)\big] \end{aligned}$$

- By optimality of θ_t: $(\Phi'(\theta_t) - \Phi'(\eta_t))^\top(\theta_t - \theta) \leqslant 0$ which is equivalent to: $D_\Phi(\theta, \eta_t) \geqslant D_\Phi(\theta, \theta_t) + D_\Phi(\theta_t, \eta_t)$. Thus

$$\begin{aligned} D_\Phi(\theta_{t-1}, \eta_t) - D_\Phi(\theta_t, \eta_t) &= \Phi(\theta_{t-1}) - \Phi(\theta_t) - \Phi'(\eta_t)^\top(\theta_{t-1} - \theta_t) \\ &\leqslant (\Phi'(\theta_{t-1}) - \Phi'(\eta_t))^\top(\theta_{t-1} - \theta_t) - \frac{\alpha}{2}\|\theta_{t-1} - \theta_t\|^2 \\ &= \gamma f'(\theta_{t-1})^\top(\theta_{t-1} - \theta_t) - \frac{\alpha}{2}\|\theta_{t-1} - \theta_t\|^2 \\ &\leqslant \gamma B\|\theta_{t-1} - \theta_t\| - \frac{\alpha}{2}\|\theta_{t-1} - \theta_t\|^2 \leqslant \frac{(\gamma B)^2}{2\alpha} \end{aligned}$$

- Thus $\sum_{u=1}^t \big[f(\theta_{t-1}) - f(\theta)\big] \leqslant \frac{D_\Phi(\theta, \theta_0)}{\gamma} + \gamma \frac{L^2 t}{2\alpha}$

Mirror descent examples

- **Euclidean**: $\Phi = \frac{1}{2}\|\cdot\|_2^2$ with $\|\cdot\| = \|\cdot\|_2$ and $\mathcal{C}_\Phi = \mathbb{R}^d$
  - Regular gradient descent
- **Simplex**: $\Phi(\theta) = \sum_{i=1}^d \theta_i \log \theta_i$ with $\|\cdot\| = \|\cdot\|_1$ and $\mathcal{C}_\Phi = \big\{\theta \in \mathbb{R}_+^d,\ \sum_{i=1}^d \theta_i = 1\big\}$
  - Bregman divergence = Kullback-Leibler divergence
  - Iteration (multiplicative update): $\theta_t \propto \theta_{t-1} \exp(-\gamma f'(\theta_{t-1}))$
  - Constant: $D^2 = \log d$, $\alpha = 1$
- ℓ_p**-ball**: $\Phi(\theta) = \frac{1}{2}\|\theta\|_p^2$, with $\|\cdot\| = \|\cdot\|_p$, $p \in (1, 2]$
  - We have $\alpha = p - 1$
  - Typically used with $p = 1 + \frac{1}{\log d}$ to cover the ℓ_1-geometry
  - See Duchi et al. (2010)

Minimax rates (Agarwal et al., 2012)

- **Model of computation (i.e., algorithms): first-order oracle**
  - Queries a function f by obtaining $f(\theta_k)$ and $f'(\theta_k)$ with zero-mean bounded variance noise, for $k = 0, \ldots, n-1$ and outputs θ_n
- **Class of functions**
  - convex B-Lipschitz-continuous (w.r.t. ℓ_2-norm) on a compact convex set $\mathcal{C}$ containing an ℓ_∞-ball
- **Performance measure**
  - for a given algorithm and function $\varepsilon_n(\text{algo}, f) = f(\theta_n) - \inf_{\theta \in \mathcal{C}} f(\theta)$
  - for a given algorithm: $\sup_{\text{functions } f} \varepsilon_n(\text{algo}, f)$
- **Minimax performance**: $\inf_{\text{algo}} \sup_{\text{functions } f} \varepsilon_n(\text{algo}, f)$

Minimax rates (Agarwal et al., 2012)

- **Convex functions**: domain $\mathcal{C}$ that contains an ℓ_∞-ball of radius D

$$\inf_{\text{algo}} \sup_{\text{functions } f} \varepsilon(\text{algo}, f) \geqslant \text{cst} \times \min\left\{ BD\sqrt{\frac{d}{n}}, BD \right\}$$

  - Consequences for ℓ_2-ball of radius D: $BD/\sqrt{n}$
  - Upper-bound through stochastic subgradient

- μ-**strongly-convex functions**:

$$\inf_{\text{algo}} \sup_{\text{functions } f} \varepsilon_n(\text{algo}, f) \geqslant \text{cst} \times \min\left\{ \frac{B^2}{\mu n}, \frac{B^2}{\mu d}, BD\sqrt{\frac{d}{n}}, BD \right\}$$

Minimax rates - sketch of proof

1. **Create a subclass of functions** indexed by some vertices α^j, $j = 1, \ldots, M$ of the hypercube $\{-1, 1\}^d$, which are sufficiently far in Hamming metric Δ_H (denote $\mathcal{V}$ this set with $|\mathcal{V}| = M$)

$$\forall j \neq k, \ \Delta_H(\alpha^i, \alpha^j) \geqslant \frac{d}{4},$$

e.g., a "$\frac{d}{4}$-packing" (possible with M exponential in d - see later)

Minimax rates - sketch of proof

1. **Create a subclass of functions** indexed by some vertices α^j, $j = 1, \ldots, M$ of the hypercube $\{-1, 1\}^d$, which are sufficiently far in Hamming metric Δ_H (denote $\mathcal{V}$ this set with $|\mathcal{V}| = M$)

$$\forall j \neq k, \ \Delta_H(\alpha^i, \alpha^j) \geqslant \frac{d}{4},$$

   e.g., a "$\frac{d}{4}$-packing" (possible with M exponential in d - see later)

2. **Design functions** so that

   – approximate optimization of the function is equivalent to function identification among the class above

   – stochastic oracle corresponds to a sequence of coin tosses with biases index by α^j, $j = 1, \ldots, M$

Minimax rates - sketch of proof

1. **Create a subclass of functions** indexed by some vertices α^j, $j = 1, \ldots, M$ of the hypercube $\{-1, 1\}^d$, which are sufficiently far in Hamming metric Δ_H (denote $\mathcal{V}$ this set with $|\mathcal{V}| = M$)

$$\forall j \neq k, \ \Delta_H(\alpha^i, \alpha^j) \geqslant \frac{d}{4},$$

   e.g., a "$\frac{d}{4}$-packing" (possible with M exponential in d - see later)

2. **Design functions** so that
   - approximate optimization of the function is equivalent to function identification among the class above
   - stochastic oracle corresponds to a sequence of coin tosses with biases index by α^j, $j = 1, \ldots, M$

3. Any such identification procedure (i.e., **a test**) has a lower bound on the probability of error

Packing number for the hyper-cube
Proof

- **Varshamov-Gilbert's lemma** (Massart, 2003, p. 105): the maximal number of points in the hypercube that are at least $d/4$-apart in Hamming loss is greater than than $\exp(d/8)$.

1. Maximality of family $\mathcal{V} \Rightarrow \bigcup_{\alpha \in \mathcal{V}} \mathcal{B}_H(\alpha, d/4) = \{-1, 1\}^d$

2. Cardinality: $\sum_{\alpha \in \mathcal{V}} |\mathcal{B}_H(\alpha, d/4)| \geqslant 2^d$

3. Link with deviation of Z distributed as $\text{Binomial}(d, 1/2)$

$$2^{-d}|\mathcal{B}_H(\alpha, d/4)| = \mathbb{P}(Z \leqslant d/4) = \mathbb{P}(Z \geqslant 3d/4)$$

4. Hoeffding inequality: $\mathbb{P}(Z - \frac{d}{2} \geqslant \frac{d}{4}) \leqslant \exp(-\frac{2(d/4)^2}{d}) = \exp(-\frac{d}{8})$

Designing a class of functions

- Given $\alpha \in \{-1, 1\}^d$, and a precision parameter $\delta > 0$:

$$g_\alpha(x) = \frac{c}{d} \sum_{i=1}^{d} \left\{ (\frac{1}{2} + \alpha_i \delta) f_i^+(x) + (\frac{1}{2} - \alpha_i \delta) f_i^-(x) \right\}$$

- **Properties**
  - Functions f_i's and constant c to ensure proper regularity and/or strong convexity

- **Oracle**

(a) Pick an index $i \in \{1, \ldots, d\}$ at random
(b) Draw $b_i \in \{0, 1\}$ from a Bernoulli with parameter $\frac{1}{2} + \alpha_i \delta$
(c) Consider $\hat{g}_\alpha(x) = c\big[b_i f_i^+ + (1 - b_i) f_i^-\big]$ and its value / gradient

Optimizing is function identification

- **Goal**: if g_α is optimized up to error ε, then this identifies $\alpha \in \mathcal{V}$
- **"Metric" between functions**:

$$\rho(f,g) = \inf_{\theta\in\mathcal{C}} f(\theta) + g(\theta) - \inf_{\theta\in\mathcal{C}} f(\theta) - \inf_{\theta\in\mathcal{C}} g(\theta)$$

  - $\rho(f,g) \geqslant 0$ with equality iff f and g have the same minimizers

- **Lemma**: let $\psi(\delta) = \min_{\alpha\neq\beta\in\mathcal{V}} \rho(g_\alpha, g_\beta)$. For any $\tilde{\theta} \in \mathcal{C}$, there is at most one function g_α such that $g_\alpha(\tilde{\theta}) - \inf_{\theta\in\mathcal{C}} g_\alpha(\theta) \leqslant \frac{\psi(\delta)}{3}$

Optimizing is function identification

- **Goal**: if g_α is optimized up to error ε, then this identifies $\alpha \in \mathcal{V}$
- **"Metric" between functions**:

$$\rho(f, g) = \inf_{\theta \in \mathcal{C}} f(\theta) + g(\theta) - \inf_{\theta \in \mathcal{C}} f(\theta) - \inf_{\theta \in \mathcal{C}} g(\theta)$$

  - $\rho(f, g) \geqslant 0$ with equality iff f and g have the same minimizers

- **Lemma**: let $\psi(\delta) = \min_{\alpha \neq \beta \in \mathcal{V}} \rho(g_\alpha, g_\beta)$. For any $\tilde{\theta} \in \mathcal{C}$, there is at most one function g_α such that $g_\alpha(\tilde{\theta}) - \inf_{\theta \in \mathcal{C}} g_\alpha(\theta) \leqslant \frac{\psi(\delta)}{3}$
  - (a) optimizing an unknown function from the class up to precision $\frac{\psi(\delta)}{3}$ leads to identification of $\alpha \in \mathcal{V}$
  - (b) If the expected minimax error rate is greater than $\frac{\psi(\delta)}{9}$, there exists a function from the set of random gradient and function values such the probability of error is less than $1/3$

Lower bounds on coin tossing (Agarwal et al., 2012, Lemma 3)

- **Lemma**: For $\delta < 1/4$, given α^* uniformly at random in $\mathcal{V}$, if n outcomes of a random single coin (out of the d) are revealed, then any test will have a probability of error greater than

$$1 - \frac{16n\delta^2 + \log 2}{\frac{d}{2}\log(2/\sqrt{e})}$$

  - Proof based on Fano's inequality: If g is a function of Y, and X takes m values, then

$$\mathbb{P}(g(X) \neq Y) \geqslant \frac{H(X|Y) - 1}{\log m} = \frac{H(X)}{\log m} - \frac{I(X,Y) + 1}{\log m}$$

Construction of f_i for convex functions

- $f_i^+(\theta) = |\theta(i) + \frac{1}{2}|$ and $f_i^-(\theta) = |\theta(i) - \frac{1}{2}|$
  - 1-Lipschitz-continuous with respect to the ℓ_2-norm. With $c = B/2$, then g_α is B-Lipschitz.
  - Calling the oracle reveals a coin

- Lower bound on the discrepancy function
  - each g_α is minimized at $\theta_\alpha = -\alpha/2$
  - Fact: $\rho(g_\alpha, g_\beta) = \frac{2c\delta}{d}\Delta_H(\alpha, \beta) \geqslant \frac{c\delta}{2} = \psi(\delta)$

- Set error/precision $\varepsilon = \frac{c\delta}{18}$ so that $\varepsilon < \psi(\delta)/9$

- Consequence: $\frac{1}{3} \geqslant 1 - \frac{16n\delta^2 + \log 2}{\frac{d}{2}\log(2/\sqrt{e})}$, that is, $\boxed{n \geqslant \text{ cst } \times \frac{L^2 d^2}{\varepsilon^2}}$

Construction of f_i for strongly-convex functions

- $f_i^{\pm}(\theta) = \frac{1}{2}\kappa|\theta(i) \pm \frac{1}{2}| + \frac{1-\kappa}{4}(\theta(i) \pm \frac{1}{2})^2$
  - Strongly convex and Lipschitz-continuous
- Same proof technique (more technical details)
- See more details by Agarwal et al. (2012); Raginsky and Rakhlin (2011)

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