Through the sheet, we would simplify $\|\cdot\|_2$ as $\|\cdot\|$ for vectors and spectral norm for matrices, if not explicitly say not so. For ease of reference in the script, we include the reference in the beginning of every referred property.

1 Introduction

A classical algorithm is evaluated on space and time complexity, but a learning algorithm is mainly evaluated on how well it explains the data.

Definitions:

1. A *risk* or *loss* is a function $l : \mathcal{H} \times \mathcal{X} \to \mathcal{R}$, representing how bad a given hypothesis $H \in \mathcal{H}$ is, on the given data X.
2. The *expected risk* is a function $l(H) := \mathbb{E}_{\mathcal{X}}(l(H, X))$, representing how bad a given hypothesis $H \in \mathcal{H}$ is, measured on the whole data distribution.
3. An optimal hypothesis is a hypothesis $H^* := \mathbf{argmin}_{H\in\mathcal{H}} l(H)$.
4. An *empirical risk minimization (ERM)* hypothesis is a hypothesis $H^*(X) := \mathbf{argmin}_{H\in\mathcal{H}} l(H, X)$.
5. The *probably approximately correct (PAC)* theory characterizes the following: under which condition that given $\delta, \epsilon > 0$, the ERM $\tilde{H} := H^*(X)$ satisfies $l(\tilde{H}) \le \inf_{H\in\mathcal{H}} l(H) + \epsilon$ with probability at least $1-\delta$. The probability is over the sampling.

Theorem: Suppose that for any $\delta, \epsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that for $n \ge n_0$, $\sup_{H\in\mathcal{H}} |\ell_n(H) - \ell(H)| \le \varepsilon$ with probability at least $1-\delta$. Then, for $n \ge n_0$, an approximate empirical risk minimizer $\tilde{H}_n$ is PAC for expected risk minimization, meaning that it satisfies $\ell\left(\tilde{H}_n\right) \le \inf_{H\in\mathcal{H}} \ell(H) + 3\varepsilon$ with probability at least $1-\delta$.

Note: the condition $\sup_{H\in\mathcal{H}} |\ell_n(H) - \ell(H)| \le \varepsilon$ is to ensure $\ell\left(\tilde{H}_n\right) \le \ell_n\left(\tilde{H}_n\right) + \varepsilon$, which cannot be proved by the law of large numbers, because it cannot be applied to $\tilde{H}_n$ which is data dependent (thus dependent to the random variables involved in the law of large numbers).

1.1 VC Theory

Setting: binary classification, unknown distribution over the data source.

For n given samples, each hypothesis $H \in \mathcal{H}$ induces a cut $H \cap \{x_1, \ldots, x_n\}$. The set of all possible cuts by the hypothesis class on n given samples is $\mathcal{H} \cap \{x_1, \ldots, x_n\} := \{H \cap \{x_1, \ldots, x_n\} : H \in \mathcal{H}\}$. For any n samples, the elements of this set is bounded by 2^n, as each sample can be either 0 or 1. Further, define $\mathcal{H}(n) := \max_{x_1,\ldots,x_n} |\mathcal{H} \cap \{x_1, \ldots, x_n\}|$.

Theorem (VC): Let $\mathcal{D}$ be a probability space, $\mathcal{H}$ a set of events, $n \in \mathbb{N}, \varepsilon > 0$. Then $\sup_{H\in\mathcal{H}} \left|\text{prob}_n(H) - \text{prob}(H)\right| > \varepsilon$ with probability at most $4\mathcal{H}(2n) \cdot \exp\left(-\varepsilon^2 n/8\right)$.

Note that in this case, $\ell(H) = \text{prob}(H)$, thus if $\mathcal{H}(n)$ is polynomially bounded, we have $\sup_{H\in\mathcal{H}} |\ell_n(H) - \ell(H)| \le \varepsilon$ for sufficiently large n, which implies PAC property.

2 Convex Functions and Optimization

Cauchy-Schwarz Inequality: Let $u, v \in \mathbb{R}^d$, then $|\langle u, v\rangle| \le \|u\|\|v\|$ for any inner product space, where $\|\cdot\|$ is the induced norm of the inner product, defined as $\|u\| = \sqrt{\langle u, u\rangle}$.

Spectral Nrom: Let A be a matrix of shape $m \times d$. The spectral norm of A is $\|A\| = \max_{\|v\|_2=1} \|Av\|_2$.

Definitions:

1. A set S is convex if for any $x, y \in S$ and any $\lambda \in [0,1]$, $\lambda x + (1-\lambda)y \in S$.
2. A function f is convex if (1) its domain is convex, (2) the function value of any convex combination is below the convex combination of the function values, i.e., for all $x, y \in \mathbf{dom}(f)$ and all $\lambda \in [0,1]$, we have $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y)$.
3. A function f is strictly convex if the inequality above is strict.
4. The epigraph of a function is the upper region characterized by the function, i.e., $\mathbf{epi}(f) = \{(x, y) : x \in \mathbf{dom}(f) \wedge y \ge f(x)\}$.

General Properties:

1. (2.11) f is convex iff $\mathbf{epi}(f)$ is a convex set. Proof by definition.
2. Jensen's Inequality: for $\lambda_i \ge 0$, $\sum_i \lambda_i = 1$ and a convex function f, then $f(\sum_i \lambda_i x_i) \le \sum_i \lambda_i f(x_i)$ for any $x_i \in \mathbf{dom}(f)$. Proof by induction.
3. (2.13) A convex function f with $\mathbf{dom}(f)$ open and $\mathbf{dom}(f) \in \mathbb{R}^d$ is continuous. Proof: first show f is bounded on any hypercube and has the extreme values on some corners. Then use sufficiently small cube to conclude continuity.
4. **First-order characterization**: A differentiable function f with open convex domain is convex iff for any $x, y \in \mathbf{dom}(f)$, we have $f(y) \ge f(x) + \nabla f(x)^\top (y-x)$. Proof: (1) One side: Taylor expansion for $f(x + t(y-x))$ on t at x and use $f(x + t(y-x)) \le f(x) + t(f(y) - f(x))$; (2) The other side: introduce $z = \lambda x + (1-\lambda)y$, then apply the inequalities $f(x) \ge f(z) + \nabla f(z)^\top (x-z)$ and $f(y) \ge f(z) + \nabla f(z)^\top (y-z)$ and plug into $\lambda f(x) + (1-\lambda) f(y)$.
5. **Monotonicity of the gradient**: A differentiable function f with open convex domain is convex iff $(\nabla f(y) - \nabla f(x))^\top (y-x) \ge 0$. Proof: (1) One side: use first-order characterization alternatively for x and y, then add these two inequalities. (2) The other side: define $h(t) = f(x + t(y-x))$, thus $h'(t) = \nabla f(x + t(y-x))^\top (y-x)$. By monotonicity of the gradient, $h'(t) \ge \nabla f(x)^\top (y-x)$ for $t \in (0,1)$. By mean-value theorem, there exists $c \in (0,1)$ such that $h(1) - h(0) = h'(c) \ge \nabla f(x)^\top (y-x)$. Plugging in $h(1) = f(y)$ and $h(0) = f(x)$ suffices.
6. **Second-order characterization**: A twice continuously differentiable function f with open convex domain is convex iff $\nabla^2 f(x) \succeq 0$. Proof: (1) One side: use $h(t) = f(x + t(y-x))$ again. By monotonicity of the gradient, we can conclude $\frac{h'(\delta)-h'(0)}{\delta} \ge 0$, which implies $h''(0) \ge 0$. Note $h''(t) = (y-x)^\top \nabla^2 f(x + t(y-x))(y-x)$, and thus $h''(0) \ge 0$ implies $\nabla^2 f(x) \succeq 0$. (2) The other side: mean-value theorem implies $h'(1) - h'(0) = h''(c)$ for some $c \in (0,1)$, which is $(\nabla f(y) - \nabla f(x))^\top (y-x) = (y-x)^\top \nabla^2 f(x + c(y-x))(y-x) \ge 0$. This is the monotonicity of the gradient, thus f is convex.
7. (2.18) **Operations that perserve convexity**: For convex functions $\{f_i\}$, $f := \max_i f_i$ and $f := \sum_i \lambda_i f_i$ for $\lambda_i \ge 0$ are convex on $\mathbf{dom}(f) := \cap_i \mathbf{dom}(f_i)$. For convex f and affine g, $f \circ g$ is also convex.

Minimizer Properties:

1. (2.20) Every local minimum of a convex function is a global minimum. Proof by contradiction, moving a small step from the local minimum towards the global minimum violates convexity.
2. (2.21 and 2.22) For a differentiable convex function with open domain, $\nabla f(x) = 0$ iff x is a global minimum. Proof: (1) One side: first-order characterization. (2) The other side: contradiction by moving a small step in the gradient descent direction.
3. (2.24) Positive definite Hessian implies strict convexity. Proof by Taylor expansion. The reverse is false, e.g., $f(x) = x^4$ is strictly convex, but its Hessian is not positive definite at 0.
4. (2.25) A stricly convex function has at most one global minimum. Proof by contradiction.
5. (2.27) For f convex and differentiable over an open domain, x^* is a minimizer over $\mathbf{dom}(f)$ iff $\nabla f(x^*)^\top (x - x^*) \ge 0$ for any $x \in \mathbf{dom}(f)$. Proof by first-order characterization.

2.1 Convex Programming

Definitions:

1. The standard form of a convex program is to minimize $f(x)$ s.t. $f_i(x) \le 0$ and $h_i(x) = 0$, where f and f_i are convex and h_i are affine. The feasible set is convex and the objective is convex.
2. The Lagrangian of a program is $L(x, \lambda, \nu) := f(x) + \sum_i \lambda_i f_i(x) + \sum_i \nu_i h_i(x)$. The λ_i and ν_i are called Lagrange multipliers.
3. The Lagrange dual function is $g(\lambda_i, \nu_i) = \inf_{x\in\mathcal{D}} L(x, \lambda_i, \nu_i)$, where $\mathcal{D}$ is the feasible set.
4. The Lagrange dual program is to maximize $g(\lambda_i, \nu_i)$ s.t. $\lambda_i \ge 0$.

Properties:

1. **Weak duality**: for any feasible x and $\lambda_i \ge 0$, we have $g(\lambda_i, \nu_i) \le f(x)$, i.e., $\max_{\lambda_i \ge 0} g(\lambda_i, \nu_i) \le \inf_{x\in\mathcal{D}} f(x)$. Proof by noticing that for feasible x, $L(x, \lambda_i, \nu_i) \le f(x)$.
2. **Strong duality (Slater's condition)** For convex program with a feasible x such that the inequalities are stricly satisfied (Slater's point), then $\max_{\lambda_i \ge 0} g(\lambda_i, \nu_i) = \inf_{x\in\mathcal{D}} f(x)$. If this value is finite, then it is attained by a feasible solution of the dual program. Slater's condition is sufficient but not necessary for strong duality.
3. **Strong duality (KKT condition)** Strong duality holds iff the followings hold for the solution (1) (complementary slackness) $\lambda_i f_i(x) = 0$, (2) (vanishing gradient) $\nabla_x L(x, \lambda_i, \nu_i) = 0$. (3) Primal feasibility and dual feasibility.

3 Gradient Descent

	Lipschitz convex functions	smooth convex functions	strongly convex functions	smooth & strongly convex functions
gradient descent	Thm. 3.1 $\mathcal{O}(1/\varepsilon^2)$	Thm. 3.8 $\mathcal{O}(1/\varepsilon)$		Thm. 3.14 $\mathcal{O}(\log(1/\varepsilon))$
accelerated gradient descent		Thm. 3.9 $\mathcal{O}(1/\sqrt{\varepsilon})$		
projected gradient descent	Thm. 4.2 $\mathcal{O}(1/\varepsilon^2)$	Thm. 4.4 $\mathcal{O}(1/\varepsilon)$		Thm. 4.5 $\mathcal{O}(\log(1/\varepsilon))$
subgradient descent	Thm. ?? $\mathcal{O}(1/\varepsilon^2)$		Thm. ?? $\mathcal{O}(1/\varepsilon)$	
stochastic gradient descent	Thm. ?? $\mathcal{O}(1/\varepsilon^2)$		Thm. ?? $\mathcal{O}(1/\varepsilon)$	

Define $g_t = \nabla f(x_t)$. The gradient descent is $x_{t+1} = x_t - \gamma g_t$, where γ is the step size. We assume f is differentiable anywhere.

Definitions:

1. A convex differentiable function f is L-smooth if $f(y) \le f(x) + \nabla f(x)^\top (y-x) + \frac{L}{2}\|x-y\|^2$ for any x, y.
2. A function is μ-strongly convex if $f(y) \ge f(x) + \nabla f(x)^\top (y-x) + \frac{\mu}{2}\|x-y\|^2$ for any x, y.

Properties:

1. **Characterization of L-smoothness (equivalent)**: (3.3) $\frac{L}{2}x^\top x - f(x)$ is convex; (3.5) $\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|$ for any x, y;
2. **Operations that perserve smoothness** (3.6): (i) Assume f_i are L_i-smooth and $\lambda_i > 0$, then $f := \sum_i \lambda_i f_i$ is $\sum_i \lambda_i L_i$-smooth. Proof by (3.3). (ii) Assume f is L-smooth and $g(x) = Ax + b$, then $f \circ g$ is $L\|A\|^2$-smooth.
3. **Characterization of μ-strongly convexity (equivalent)**: (3.11) $f(x) - \frac{\mu}{2}x^\top x$ is convex.
4. **Bound of first-order changes**: Let f be convex and x^* be the global minimum, i.e. $\nabla f(x^*) = 0$. If f is μ-strongly convex, then $\nabla f(x)^\top (x - x^*) \ge \mu\|x - x^*\|^2$. If f is L-smooth, then $\nabla f(x)^\top (x-x^*) \le L\|x-x^*\|^2$. Proof: write definition separately for x^* on x_t and x_t on x^*, then add the two inequalities together.

Analysis based on the first-order characterization:

1. Vanilla analysis (no further assumption): by first-order characterization, we can bound $f(x_t) - f(x^*) \le g_t^\top (x_t - x^*)$. The algorithm says $g_t = (x_t - x_{t+1})/\gamma$, thus $g_t^\top (x_t - x^*) = \frac{1}{\gamma}(x_t - x_{t+1})^\top (x_t - x^*)$. Since $2a^\top b = a^\top a + b^\top b - (a-b)^\top (a-b)$, we have $2(x_t - x_{t+1})^\top (x_t - x^*) = \|x_t - x_{t+1}\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2$. Therefore, $\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \le \sum_{t=0}^{T-1} g_t^\top (x_t - x^*) \le \frac{\gamma}{2}\sum_{t=0}^{T-1}\|g_t\|^2 + \frac{1}{2\gamma}\|x_0 - x^*\|^2$. The problem is to bound the squared norm of gradients.
2. Lipschitz convex functions: bounded gradients $\|\nabla f(x)\| \le B$ for any x. (3.1) Assume $\|x_0 - x^*\| \le R$. Then the result of vanilla analysis says $\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \le \frac{\gamma}{2}B^2 T + \frac{1}{2\gamma}R^2$. Choose $\gamma = \frac{R}{B\sqrt{T}}$ yields $\frac{1}{T}\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \le \frac{RB}{\sqrt{T}}$. This means we need $T \ge R^2B^2/\epsilon^2$ to achieve $\min_t (f(x_t) - f(x^*)) \le \epsilon$.

Analysis based on control over the quadratic term:

1. L-smooth functions (not requiring convexity): (3.7) with $\gamma := 1/L$, we have $f(x_{t+1}) \le f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|^2$. Proof: use smoothness definition and plug in $x_{t+1} - x_t = -\frac{1}{L}\nabla f(x_t)$.
2. Convex L-smooth functions: (3.8) with $\gamma := 1/L$, we have $f(x_T) - f(x^*) \le \frac{L}{2T}\|x_0 - x^*\|^2$. This means we only need $T \ge \frac{R^2 L}{2\epsilon}$ to achieve error at most ϵ. Proof: by (3.7), $\frac{1}{2L}\sum_{t=0}^{T-1}\|\nabla f(x_t)\|^2 \le f(x_0) - f(x^*)$. Plugging into vanilla analysis, we have $\sum_{t=1}^{T}(f(x_t) - f(x^*)) \le \frac{L}{2}\|x_0 - x^*\|^2$. By (3.7), $f(x_t) - f(x^*)$ is monotonically decreasing, thus $f(x_T) - f(x^*) \le \frac{L}{2T}\|x_0 - x^*\|^2$.
3. μ-strongly convex and L-smooth functions: (3.14) GD with $\gamma = 1/L$ yields $\|x_{t+1} - x^*\|^2 \le (1 - \frac{\mu}{L})\|x_t - x^*\|^2$ and $f(x_T) - f(x^*) \le \frac{L}{2}(1 - \frac{\mu}{L})^T\|x_0 - x^*\|^2$. This means we need $T \ge \frac{L}{\mu}\log\left(\frac{R^2 L}{2\epsilon}\right)$ to achieve error at most ϵ. Proof: (i) replacing the first-order characterization in the vanilla analysis by the condition of μ-strongly convexity, we get $\|x_{t+1} - x^*\|^2 \le 2\gamma(f(x^*) - f(x_t)) + \gamma^2\|\nabla f(x_t)\|^2 + (1 - \mu\gamma)\|x_t - x^*\|^2$. By sufficient decrease of L-smooth functions, we have $f(x^*) - f(x_t) \le f(x_{t+1}) - f(x_t) \le -\frac{1}{2L}\|\nabla f(x_t)\|^2$. Combining these two gives the first result. (ii) By smoothness, $f(x_T) - f(x^*) \le \frac{L}{2}\|x_T - x^*\|^2 \le \frac{L}{2}(1 - \frac{\mu}{L})^T\|x_0 - x^*\|^2$.

Optimizing without knowing L or B: for L-smooth convex functions, we do not need to know L to ensure $O(\frac{R^2 L}{\epsilon})$ steps. The idea is to guess L and refine it gradually. The first guess is $L_0 := \frac{2\epsilon}{R^2}$. For each guess, we check if the sufficient decrease (3.7) holds. If (3.7) holds in the whole process $T = \frac{R^2 L_i}{2\epsilon}$, then L_i is a successful guess, and we finish. If the guess is incorrect, we double $L_{i+1} = 2L_i$ and repeat. The final guess cannot exceed two times the correct value, and the number of iterations is bounded by $\sum_i 2^i$ until the last term exceeds the true L. This means the total number of iterations is bounded by two times the last iteration time, which is still $O(\frac{R^2 L}{\epsilon})$. A similar approach can be taken for Lipschitz convex functions to optimize without knowing B.

Accelerated Gradient Descent

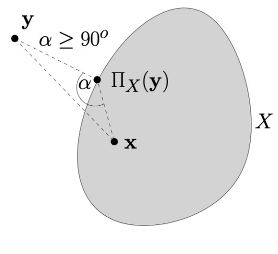
The AGD is to do $y_{t+1} = x_t - \frac{1}{L}\nabla f(x_t)$, $z_{t+1} = z_t - \frac{t+1}{2L}\nabla f(x_t)$ and $x_{t+1} = \frac{t+1}{t+3}y_{t+1} + \frac{2}{t+3}z_{t+1}$. Initialized with $y_0 = z_0 = x_0$.

(Thm 3.9) For L-smooth convex functions, AGD yields $f(y_T) - f(x^*) \le \frac{2L\|z_0 - x^*\|^2}{T(T+1)}$.

4 Projected Gradient Descent

PGD: do constrainted gradient descent. $y_{t+1} = x_t - \gamma\nabla f(x_t)$ then $x_{t+1} = \Pi_X(y_{t+1}) = \mathbf{argmin}_X \|x - y_{t+1}\|^2$.

projection inequalities (4.1): let X be closed and convex, $x \in X$, then $(x - \Pi_X(y))^\top (y - \Pi_X(y)) \le 0$ and $\|x - \Pi_X(y)\|^2 + \|y - \Pi_X(y)\|^2 \le \|x-y\|^2$. Proof: $\Pi_X(y)$ minimizes $f(x) = \|x - y\|^2$ on Z, thus by optimality, $\nabla f(\Pi_X(y))^\top (x - \Pi_X(y)) = 2(\Pi_X(y) - y)^\top (x - \Pi_X(y)) \ge 0$ for any $x \in X$. The second follows from $2a^\top b = \|a\|^2 + \|b\|^2 - \|a-b\|^2$. Illustration:



Analysis for PGD: the same result but proof adapted using projection inequalities.

1. Vanilla analysis: the same analysis for PGD gives $g_t^\top (x_t - x^*) = \frac{1}{2\gamma}\left(\gamma^2\|g_t\|^2 + \|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2\right)$. By projection inequality, we have $\|x_{t+1} - x^*\|^2 \le \|y_{t+1} - x^*\|^2$. Thus, the result of vanilla analysis is the same.
2. Lipschitz convex functions: (4.2) the same as GD since it only requires the result of vanilla analysis.
3. L-smooth functions: (4.3) with $\gamma = 1/L$, we have $f(x_{t+1}) \le f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|^2 + \frac{L}{2}\|y_{t+1} - x_{t+1}\|^2$. In addition, $f(x_{t+1}) \le f(x_t)$. Proof: write the smoothness definition for x_{t+1} for x_t. Replace $\nabla f(x_t)^\top (x_{t+1} - x_t)$ by $-L(y_{t+1} - x_t)^\top (x_{t+1} - x_t)$ and then apply $2a^\top b = \|a\|^2 + \|b\|^2 - \|a-b\|^2$ gives the first result. Since $\|y_{t+1} - x_{t+1}\| \le \|y_{t+1} - x_t\| = \gamma\|\nabla f(x_t)\|$, we have $f(x_{t+1}) \le f(x_t)$.
4. Convex L-smooth functions: (4.4) the same result as GD. Proof: we use a tighter inequality for vanilla analysis. Instead of using $\|x_{t+1} - x^*\|^2 \le \|y_{t+1} - x^*\|^2$, we now use $\|x_{t+1} - x^*\|^2 + \|y_{t+1} - x_{t+1}\|^2 \le \|y_{t+1} - x^*\|^2$, so the vanilla analysis results in $\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \le \frac{1}{2L}\sum_{t=0}^{T-1}\|g_t\|^2 + \frac{L}{2}\|x_0 - x^*\|^2 - \frac{L}{2}\sum_{t=0}^{T-1}\|y_{t+1} - x_{t+1}\|^2$. Combining this with (4.3) gets the result.
5. μ-strongly convex and L-smooth functions: (4.5) PGD with $\gamma = 1/L$ yields $\|x_{t+1} - x^*\|^2 \le (1 - \frac{\mu}{L})\|x_t - x^*\|^2$ and $f(x_T) - f(x^*) \le \frac{L}{2}(1 - \frac{\mu}{L})^T\|x_0 - x^*\|^2 + (1 - \frac{\mu}{L})^{T/2}\|\nabla f(x^*)\|\|x_0 - x^*\|$. This is still $O(\log(\frac{1}{\epsilon}))$ steps. Proof: μ-strongly convexity strengthens the vanilla analysis to be $g_t^\top (x_t - x^*) \le \frac{1}{2\gamma}(\gamma^2\|g_t\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2 - \|y_{t+1} - x_{t+1}\|^2) - \frac{\mu}{2}\|x_t - x^*\|^2$. This makes the vanilla analysis to give $\|x_{t+1} - x^*\|^2 \le 2\gamma(f(x^*) - f(x_t)) + \gamma^2\|\nabla f(x_t)\|^2 + (1 - \mu\gamma)\|x_t - x^*\|^2 - \|y_{t+1} - x_{t+1}\|^2$. The extra $-\|y_{t+1} - x_{t+1}\|^2$ happens to compensate for the additional term from the L-smoothness, thus (i) follows. By smoothness, we have $f(x_T) - f(x^*) \le \|\nabla f(x^*)\|\|x_T - x^*\| + \frac{L}{2}\|x_T - x^*\|^2$, thus (ii) follows from (i).

Projecting into L_1 ball: solve $\Pi_X(v) = \mathbf{argmin}_{\|x\| \le 1}\|x - v\|^2$. WLOG, assume $v_1 \ge v_2 \ge \dots v_d \ge 0$ and $\sum_i v_i > 1$. (4.11) We have $x_i^* = v_i - \theta_p$ for $i \le p$ and $x^* = 0$ for $i > p$, where $\theta_p = \frac{1}{p}(\sum_{i=1}^{p} v_i - 1)$ and $p = \max\{p \in [d] : v_p - \frac{1}{p}(\sum_{i=1}^{p} v_i - 1) > 0\}$. This makes the projection $O(d\log d)$. Actually it can be improved to $O(d)$.

5 PL Condition and Coordinate Descent

Algorithm	PL norm	Smoothness	Bound	Result
Randomized	ℓ_2	L	$1-\frac{\mu}{dL}$	Thm. 5.6
Importance sampling	ℓ_2	$(L_1, L_2, \ldots, L_d)$	$1-\frac{\mu}{d\bar{L}}$	Thm. 5.7
Steepest	ℓ_2	L	$1-\frac{\mu}{dL}$	Cor. 5.8
Steeper (than Steepest)	ℓ_1	L	$1-\frac{\mu_1}{L}$	Thm. 5.10

5.1 PL Condition

Definitions:

1. **Polyak-Lojasiewicz condition**: we say f satisfies PL condition if $\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x)-f(x^*))$.
2. **Strong convexity w.r.t.** $\|\cdot\|_1$: $f(y) \geq f(x) + \nabla f(x)^\top(y-x) + \frac{\mu}{2}\|y-x\|_1^2$.

Properties:

1. **PL condition is weaker than strong convexity**: (5.2) if f is μ-strongly convex, then f satisfies PL condition for the same μ. Proof: by strong convexity, $f(x^*)-f(x) \geq \|\nabla f(x)\|\|y-x\| + \frac{\mu}{2}\|y-x\|^2 \geq -\frac{1}{2\mu}\|\nabla f(x)\|^2$.
2. **PL condition is strictly weaker than strong convexity**: $f(x_1, x_2) = x_1^2$ is not strongly convex, but satisfies PL condition.
3. **Smooth functions satisfying PL condition can be solved by GD in** $O(\log(\frac{1}{\epsilon}))$: for such functions, we have $f(x_T)-f(x^*) \leq (1-\frac{\mu}{L})^T(f(x_0)-f(x^*))$. Proof: by sufficient descrease of L-smoothness, we have $f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|^2$. By PL condition, this means $f(x_{t+1}) \leq f(x_t) - \frac{\mu}{L}(f(x_t)-f(x^*))$, which implies $f(x_{t+1})-f(x^*) \leq (1-\frac{\mu}{L})(f(x_t)-f(x^*))$.
4. **Strong convexity w.r.t.** $\|\cdot\|_1$ **and** $\|\cdot\|_2$: if f is μ-strongly convex w.r.t. $\|\cdot\|_1$, then f is also μ-strongly convex w.r.t. $\|\cdot\|_2$; if f is μ-strongly convex w.r.t. $\|\cdot\|_2$, then f is μ/d-strongly convex w.r.t. $\|\cdot\|_1$. Proof: use $\|x\|_1 \geq \|x\|_2$ and $\|x\|_2 \geq \frac{1}{\sqrt{d}}\|x\|_1$.
5. **Strong convexity w.r.t.** $\|\cdot\|_1$ **implies PL condition w.r.t.** $\|\cdot\|_\infty$: (5.9) if f has strong convexity w.r.t. $\|\cdot\|_1$, then $\frac{1}{2}\|\nabla f(x)\|_\infty^2 \geq \mu(f(x)-f(x^*))$. Proof: simply use $\nabla f(x)^\top(y-x) \leq \|\nabla f(x)\|_\infty\|y-x\|_1$ in (5.2) instead.

5.2 Coordinate Descent

Coordinate-wise smoothness: f is called coordinate-wise smooth with parameter $L = (L_1, \ldots, L_d)$ if for every coordinate i we have $f(x+\lambda e_i) \leq f(x) + \lambda\nabla_i f(x) + \frac{L_i}{2}\lambda^2$.

Coordinate Descent Algorithm: for each iteration t, choose an active coordinate $i \in [d]$, then do $x_{t+1} = x_t - \gamma_i \nabla_i f(x_i) e_i$.

Sufficient decrease: (5.5) with $\gamma_i = 1/L_i$, we have $f(x_{i+1}) \leq f(x_i) - \frac{1}{2L_i}|\nabla_i f(x_t)|^2$. Proof: plugging the update step into the definition of coordinate-wise smoothness immediately gives the result.

Analysis for coordinate-wise smooth functions satisfying PL condition:

1. **Randomized coordinate descent**: (5.6) choose coordinate uniformly at random and set $\gamma_i = 1/L$ where $L = \max_i L_i$, the randomized coordinate descent gets $\mathbb{E}(f(x_T)-f(x^*)) \leq (1-\frac{\mu}{dL})^T(f(x_0)-f(x^*))$. This means randomized coordinate descent is the same good as GD, as the number of iterations is d times higher, but each iteration is d times cheaper. Proof: by sufficient decrease of coordinate descent, $\mathbb{E}(f(x_{t+1}) \mid x_t) \leq f(x_t) - \frac{1}{2L}\sum_{i=1}^d \frac{1}{d}|\nabla_i f(x_t)|^2 = f(x_t) - \frac{1}{2dL}\|\nabla f(x_t)\|^2$. By PL condition, this means $\mathbb{E}(f(x_{t+1}) \mid x_t) \leq f(x_t) - \frac{\mu}{dL}(f(x_t)-f(x^*))$, which implies $\mathbb{E}(f(x_{t+1})-f(x^*) \mid x_t) \leq (1-\frac{\mu}{dL})(f(x_t)-f(x^*))$. This means $\mathbb{E}(f(x_{t+1})-f(x^*)) \leq (1-\frac{\mu}{dL})\mathbb{E}(f(x_t)-f(x^*))$.
2. **Importance sampling**: (5.7) choose coordinate i with probability $\frac{L_i}{\sum_j L_j}$ and define $\bar{L} = \frac{1}{d}\sum_{i=1}^d L_i$, we have $\mathbb{E}(f(x_T)-f(x^*)) \leq (1-\frac{\mu}{d\bar{L}})^T(f(x_0)-f(x^*))$. Note how randomized coordinate descent for L-smooth functions is a special case of this result. Proof: $\mathbb{E}(f(x_{t+1}) \mid x_t) \leq f(x_t) - \frac{1}{2}\sum_{i=1}^d \frac{L_i}{\sum_j L_j}\frac{1}{L_i}|\nabla_i f(x_t)|^2 = f(x_t) - \frac{1}{2d\bar{L}}\sum_i|\nabla_i f(x_t)|^2$. The rest is the same to (5.6).
3. **Steepest coordinate descent**: (5.8) choose coordinate $i = \mathbf{argmax}_i|\nabla_i f(x_t)|$, we have $f(x_T)-f(x^*) \leq (1-\frac{\mu}{dL})^T(f(x_0)-f(x^*))$. Proof: simply remove the expectation and use maximum is greater than average.
4. **Steepest descent for strong convexity w.r.t.** $\|\cdot\|_1$: (5.10) if f is μ_1-strongly convex w.r.t. $\|\cdot\|_1$, then with $\gamma_i = 1/L$, steepest descent gives $f(x_T)-f(x^*) \leq (1-\frac{\mu_1}{L})^T(f(x_0)-f(x^*))$. Note that steepest descent has the same per iteration cost as GD, but the μ_1 could be greater than the L_2 strong convexity parameter. Proof: by sufficient decrease and the update rule, $f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|_\infty^2$. By the PL condition w.r.t. $\|\cdot\|_\infty$, this means $f(x_{t+1}) \leq f(x_t) - \frac{\mu_1}{L}(f(x_t)-f(x^*))$ and thus $f(x_{t+1})-f(x^*) \leq (1-\frac{\mu_1}{L})(f(x_t)-f(x^*))$.
5. **Greedy coordinate descent**: choose some coordinate, then do $x_{t+1} = \mathbf{argmin}_\lambda f(x_t + \lambda e_i)$, i.e., make the largest step possible. For differentiable convex functions, as the update can only be better than before, it does not compromise the analysis. For non-differentiable case, however, it may get stuck in non-optimal points. (5.11) Assume $f(x) = g(x)+h(x)$, where $h(x) = \sum_i h_i(x_i)$, g is convex and differentiable and h_i is convex, then if the greedy descent converges, it converges to the global minimum of f. Such h is called *separable*, which includes L_1 norm and squared L_2 norm. This means LASSO and ridge objectives are concrete cases. Proof: f is the sum of convex functions, thus convex. Let x be the converged point and y be a near point. Thus, $\nabla_i g(x)(y_i - x_i) + h_i(y_i) - h_i(x_i) \geq 0$, since the algorithm converges. By first-order characterization, we have $f(y)-f(x) \geq \nabla g(x)^\top(y-x) + \sum_i(h_i(y_i) - h_i(x_i)) \geq 0$.

6 Subgradient Methods

Separating plane of two convex sets: let S and T be two nonempty convex sets. A hyperplane $a^\top x = b$ is said to separate S and T if $S \cup T \not\subset H$, $S \subset H^- = \{x : a^\top x \leq b\}$ and $T \subset H^+ = \{x : a^\top x \geq b\}$. If further $S \subset H^{--} = \{x : a^\top x < b\}$ and $T \subset H^{++} = \{x : a^\top x > b\}$, then H is said to strictly separate S and T.

Hyperplane separation theorem: for two non-empty convex sets S and T, they can be separated by a hyperplane iff $\mathbf{rint}(S) \cap \mathbf{rint}(T) = \emptyset$.

Definitions:

1. **Subgradient**: let f be a convex function, then a vector g is a subgradient of f at x if $f(y) \geq f(x) + g^\top(y-x)$.
2. **Subdifferential**: the set of all subgradient at x is called the subdifferential of f at x, denoted as ∂f.
3. **Directional derivative**: the directional derivative of f at x along d is $f'(x;d) = \lim_{\delta\to 0^+}\frac{f(x+\delta d)-f(x)}{\delta}$.

Properties:

1. **subgradient = gradient, when convex and differentiable**: (6.2) if f is convex and differentiable at x, then $\partial f = \{\nabla f(x)\}$; (6.3) if f is only differentiable but not convex, then $\partial f \subseteq \{\nabla f(x)\}$. Proof: clearly $\{\nabla f(x)\} \subseteq \partial f$ given f convex; assume $\partial f \neq \emptyset$, let $y = x + \epsilon d$ for small ϵ, then $\frac{f(y)-f(x)}{\epsilon} \geq g^\top d$, which means $\nabla f(x)^\top d \geq g^\top d$ for any d. This implies $g = \nabla f(x)$.
2. **Bounded subgradient for convex functions = Lipschitz**: (6.5) if f is convex and $\mathbf{dom}(f)$ is open, then $\|g\| \leq B$ for all x and $g \in \partial f(x)$ is equivalent to $|f(x)-f(y)| \leq B\|x-y\|$ for all x,y. Proof: (i) one side: let $y = x + \epsilon g$ for some $\epsilon > 0$. Then $f(y)-f(x) \geq \epsilon\|g\|^2$. By B-Lipschitz, $f(y)-f(x) \leq B\|y-x\| = \epsilon B\|g\|$. These two imply $\|g\| \leq B$. (ii) the other side: for any x,y and $g \in \partial f(x)$, $f(x)-f(y) \leq g^\top(x-y) \leq \|g\|\|x-y\| \leq B\|x-y\|$ and $f(y)-f(x) \geq g^\top(y-x) \geq -B\|y-x\|$.
3. **Convexity means subgradient almost everywhere**: (6.10) let f be convex and $x \in \mathbf{rint}(\mathbf{dom}(f))$, then $\partial f(x)$ is non-empty and bounded. Proof: (i) non-emptiness: w.l.o.g, assume $\mathbf{dom}(f)$ is full-dimensional and $x \in \mathrm{int}(\mathbf{dom}(f))$. Since $\mathbf{epi}(f)$ is convex, by hyperplane separation theorem, $\exists s, \beta$, s.t. $s^\top y + \beta t \geq s^\top x + \beta f(x)$ for any $(y,t) \in \mathbf{epi}(f)$. Since t can be arbitratily large, we have $\beta \geq 0$. If $\beta = 0$, then $s^\top(y-x) \geq 0$ for any y, which means $s = 0$, a contradiction. Thus, $\beta > 0$. Setting $g = -\beta^{-1}s$, we get $f(y) \geq f(x) + g^\top(y-x)$. (ii) boundness: suppose $\exists g_k \in \partial f(x)$ s.t. $\|g_k\| \to +\infty$. Since $x \in \mathrm{int}(\mathbf{dom}(f))$, $\exists\delta > 0$, s.t. $B(x,\delta) \subseteq \mathbf{dom}(f)$. Therefore, $y_k := x + \delta\frac{g_k}{\|g_k\|} \in \mathbf{dom}(f)$. By definition, $f(y_k) \geq f(x) + g_k^\top(y_k - x) = f(x) + \delta\|g_k\| \to +\infty$, a contradition.
4. **Subgradient everywhere means convexity**: if $\mathbf{dom}(f)$ is convex and for any $x \in \mathbf{dom}(f)$, $\partial f(x)$ is non-empty, then f is convex. Proof: for any $x,y \in \mathbf{dom}(f)$ and $\lambda \in (0,1)$, let $z = \lambda x + (1-\lambda)y$ and $g \in \partial f(z)$. Then $f(x) \geq f(z) + g^\top(x-z)$ and $f(y) \geq f(z) + g^\top(y-z)$. Adding them leads to definition of convexity.
5. $\mathbf{dist}(0, \partial f(x))$ **decides optimality**: if $0 \in \partial f(x)$, then x is a global minimum. Proof:

by definition.

6. **Subgradient and directional derivative**: (6.13) let f be convex and $x \in \text{int}(\textbf{dom}(f))$, then $f'(x;d) = \max_{g \in \partial f(x)} g^\top d$. Proof: by the definition of subgradient. we have $f(x+\delta d) - f(x) \geq \delta g^\top d$, which implies $f'(x;d) \geq g^\top d$ for any $g \in \partial f(x)$ and thus $f'(x;d) \geq \max_{g \in \partial f(x)} g^\top d$. To conclude the other side, consider $C_1 = \{(y,t) : f(y) < t\}$ and $C_2 = \{(y,t) : y = x + \alpha d, t = f(x) + \alpha f'(x;d), \alpha \geq 0\}$. Clearly they are convex and non-empty. If $C_1 \cap C_2 = \emptyset$, then by hyperplane separation theorem, $\exists g_0, \beta$, s.t. $g_0^\top(x+\alpha d) + \beta(f(x) + \alpha f'(x;d)) \leq g_0^\top y + \beta t$ for any $\alpha \geq 0$ and any $t > f(y)$. Similar to the proof of (6.10), we can show $\beta > 0$. Let $\tilde{g} = \beta^{-1} g_0$, we have $\tilde{g}^\top(x+\alpha d) + f(x) + \alpha f'(x;d) \leq \tilde{g}^\top y + f(y)$. Set $\alpha = 0$, we have $\tilde{g}x + f(x) \leq \tilde{g}^\top y + f(y)$, which means $-\tilde{g} \in \partial f(x)$. Further, set $y = x$ and $\alpha = 1$, we have $f'(x;d) \leq -\tilde{g}^\top d \leq \max_{g \in \partial f(x)} g^\top d$.

Calculus of subgradient:

1. For $h(x) = \lambda f(x) + \mu g(x)$ where $\lambda, \mu \geq 0$ and f, g both are convex, then $\partial h(x) = \lambda \partial f(x) + \mu \partial g(x)$.
2. For $h(x) = f(Ax+b)$ where f is convex, then $\partial h(x) = A^\top \partial f(Ax+b)$.
3. For $h(x) = \sup_{\alpha \in A} f_\alpha(x)$ and each $f_\alpha(x)$ is convex, then $\partial h(x) \supseteq \text{conv}\{\partial f_\alpha(x) : f_\alpha(x) = h(x)\}$.
4. For $h(x) = F(f_1(x), \ldots, f_m(x))$ where F is non-decreasing and convex, then $\partial h(x) \supseteq \{\sum_{i=1}^m d_i \partial f_i(x) : (d_1, \ldots, d_m) \in \partial F(y_1, \ldots, y_m)\}$.

6.1 Subgradient descent

Consider f convex (possibly non-differentiable) on a closed and convex set X. The subgradient descent does $x_{t+1} = \Pi_X(x_t - \gamma_t g(x_t))$, where $g(x_t) \in \partial f(x_t)$. When f is differentiable, this reduces to PGD. However, moving towards the negative direction of subgradient is not necessarily decreasing the objective. Therefore, we can only measure $\min_{t \leq T} f(x_t) - f^*$ instead of $f(x_T) - f^*$.

Setting	Algorithm	Convex	Strongly Convex
Nonsmooth	**Subgradient method**	$O\left(\frac{B \cdot R}{\sqrt{t}}\right)$	$O\left(\frac{B^2}{\mu t}\right)$
Smooth	**Gradient descent**	$O\left(\frac{L \cdot R^2}{t}\right)$	$O\left(\left(1 - \frac{\mu}{L}\right)^t\right)$
	Accelerated gradient descent	$O\left(\frac{L \cdot R^2}{t^2}\right)$	$O\left(\left(1 - \sqrt{\frac{\mu}{L}}\right)^t\right)$

Analysis for convex functions:

1. **General analysis**: (6.17) starting from x_1, we have $\min_{1 \leq t \leq T} f(x_t) - f^* \leq \frac{1}{2}(\sum_{t=1}^T \gamma_t)^{-1}(\|x_1 - x^*\|^2 + \sum_{t=1}^T \gamma_t^2 \|g(x_t)\|^2)$ and $f(\hat{x}_T) - f^* \leq \frac{1}{2}(\sum_{t=1}^T \gamma_t)^{-1}(\|x_1 - x^*\|^2 + \sum_{t=1}^T \gamma_t^2 \|g(x_t)\|^2)$ for $\hat{x}_t = (\sum_{t=1}^T \gamma_t)^{-1}(\sum_{t=1}^T \gamma_t x_t)$. Proof: by definition, $\|x_{t+1} - x^*\|^2 = \|\Pi_X(x_t - \gamma_t g(x_t)) - x^*\|^2 \leq \|x_t - \gamma_t g(x_t) - x^*\|^2 = \|x_t - x^*\|^2 - 2\gamma_t g(x_t)^\top(x_t - x^*) + \gamma_t^2 \|g(x_t)\|^2$. By definition of subgradient, we have $f(x_t) - f^* \leq g^\top(x_t - x^*)$. These two gives $\sum_{t=1}^T \gamma_t(f(x_t) - f^*) \leq \frac{1}{2}(\|x_1 - x^*\|^2 - \|x_{T+1} - x^*\|^2 + \sum_{t=1}^T \gamma_t^2 \|g(x_t)\|^2) \leq \frac{1}{2}(\|x_1 - x^*\|^2 + \sum_{t=1}^T \gamma_t^2 \|g(x_t)\|^2)$. In addition, we have $\min_t f(x_t) - f^* \leq (\sum_{t=1}^T \gamma_t)^{-1}(\sum_t \gamma_t(f(x_t) - f^*))$, thus the first claim follows. For the second claim, use $\sum_t \gamma_t(f(x_t) - f^*) \geq (\sum_t \gamma_t)(f(\hat{x}_t) - f^*)$ by convexity.
2. **Lipschitz functions**: assume $\|x_1 - x^*\| \leq R$ and $\|g(x_t)\| \leq B$, then $\min f(x_t) - f^* \leq \frac{1}{2}(\sum_{t=1}^T \gamma_t)^{-1}(R^2 + \sum_t \gamma_t^2 B^2)$.

$O(1/\sqrt{t})$ **convergence under different stepsizes for convex Lipschitz functions**:

1. $\gamma_t = \gamma$: $\epsilon_t = \frac{1}{2}(\frac{R^2}{T\gamma} + B^2\gamma) \to \frac{B^2}{2}\gamma$. Choose $\gamma_t = \frac{R}{B\sqrt{T}}$, we have $\epsilon_t \leq \frac{RB}{\sqrt{T}}$.
2. $\sum_t \gamma_t \to +\infty$ and $\gamma_t \to 0$: we have $\epsilon_t \to 0$. If set $\gamma_t = \frac{R}{B\sqrt{t}}$, we get $\epsilon_t = O(\frac{BR}{\sqrt{T}})$
3. $\sum_t \gamma_t \to +\infty$ and $\sum_t \gamma_t^2 < +\infty$: $\epsilon_t \to 0$.
4. Polyak stepsize $\gamma_t = \frac{f(x_t) - f^*}{\|g(x_t)\|^2}$: this makes the general analysis to give $\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 - 2\gamma_t g(x_t)^\top(x_t - x^*) + \gamma_t^2 \|g(x_t)\|^2 \leq \|x_t - x^*\|^2 - 2\gamma_t(f(x_t) - f^*) + \gamma_t^2 \|g(x_t)\|^2 = \|x_t - x^*\|^2 - \frac{(f(x_t) - f^*)^2}{\|g(x_t)\|^2} \leq \|x_t - x^*\|^2 - \frac{(f(x_t) - f^*)^2}{B^2}$, thus guarantees the decrease of $\|x_t - x^*\|$. This implies $\sum_t (f(x_t) - f^*)^2 \leq R^2 B^2$, thus $\epsilon_t = O(\frac{RB}{\sqrt{T}})$.

$O(1/t)$ **convergence for strongly convex Lipschitz functions**:

1. (6.18) Let f be μ-strongly convex. With $\gamma_t = \frac{1}{\mu t}$, we have $\min f(x_t) - f^* \leq \frac{B^2(\log(T)+1)}{2\mu T}$ and $f(\hat{x}_T) - f^* \leq \frac{B^2(\log(T)+1)}{2\mu T}$ for $\hat{x}_T = \frac{1}{T}\sum_t x_t$. Proof: by strong convexity, $f(y) - f(x) + g(x)^\top(y - x) + \frac{\mu}{2}\|y - x\|^2$. Thus, general analysis gives $\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 - 2\gamma_t(f(x_t) - f^* + \frac{\mu}{2}\|x_t - x^*\|^2) + \gamma_t^2 \|g(x_t)\|^2$, which implies an upper bound on $f(x_t) - f^*$. Following the same steps in the general analysis, we get the desired result.
2. (6.19) Let f be μ-strongly convex. With $\gamma_t = \frac{2}{\mu(t+1)}$, we have $\min f(x_t) - f^* \leq \frac{2B^2}{\mu(T+1)}$ and $f(\hat{x}_T) - f^* \leq \frac{2B^2}{\mu(T+1)}$ for $\hat{x}_T = \sum_t \frac{2t}{T(T+1)} x_t$. Proof: the same analysis gives $t(f(x_t) - f^*) \leq \frac{\mu t(t-1)}{4}\|x_t - x^*\|^2 - \frac{\mu t(t+1)}{4}\|x_{t+1} - x^*\|^2 + \frac{B^2}{\mu(t+1)}$. The result follows.

Subgradient descent is asymptotically optimal for first-order subgradient methods: (Thm 6.20): for any $1 \leq t \leq n$ and $x_1 \in \mathbb{R}^n$, there exists a B-Lipschitz continuous function f and a convex set X with diameter R, s.t. for any first-order method that generates $x_t \in x_1 + \text{span}(g(x_1), \ldots, g(x_{t-1}))$, where $g(x_i) \in \partial f(x_i)$, we have $\min_{1 \leq s \leq t} f(x_s) - f^* \geq \frac{BR}{4(1+\sqrt{t})}$. In addition, there exists a B-Lipschitz continuous function f that is μ-strongly convex, s.t. $\min_{1 \leq s \leq t} f(x_s) - f^* \geq \frac{B^2}{8\mu t}$. Proof: W.l.o.g., assume $x_1 = 0$. Let $X = \{x : \|x\| \leq R/2\}$ and $f(x) = C\max_{1 \leq i \leq t} x_i + \frac{\mu}{2}\|x\|^2$ for some $C > 0, \mu > 0$. The subgradient of f is $\partial f(x) = \mu x + C \cdot \text{conv}\{e_i : i \text{ s.t. } x_i = \max_{1 \leq j \leq t} x_j\}$. The optima of f is $f^* = -\frac{C^2}{2\mu t}$. Consider $g(x) = Ce_i + \mu x$, where i is the first coordinate that $x_i = \max_{1 \leq j \leq t} x_j$. Since the algorithm runs for t iterations, the last dimension cannot be updated. Therefore, $\min_{1 \leq s \leq t} f(x_s) - f^* \geq \frac{C^2}{2\mu t}$. (1) Let $C = \frac{B\sqrt{t}}{1+\sqrt{t}}$ and $\mu = \frac{2B}{R(1+\sqrt{t})}$, we have $\max_{g \in \partial f(x)} \|g\| \leq C + \mu\|x\| \leq R$, and the first result follows. (2) Let $C = \frac{B}{2}$ and $\mu = \frac{B}{R}$, we have $\max_{g \in \partial f(x)} \|g\| \leq C + \mu\|x\| \leq R$, f is μ-strongly convex, and the second result follows.

6.2 Mirror Descent

Mirror descent only uses subgradients, and have the same aymptotic performance as subgradient descent. However, it may have better constants than subgradient descent. Subgradient descent is a special case of mirror descent.

Definitions:

1. **Bregman Divergence**: let $w(x)$ be strictly convex and continuously differentiable on a close convex X. $V_w(x,y) = w(x) - w(y) - \nabla w(y)^T(x-y)$ is defined to be the Bregman divergence. This is asymmetric, thus not a valid distance. If w is μ-strongly convex, then $V_w(x,y)$ is μ-strongly convex in x.
2. **Prox-mapping**: given an input x and vector ξ, the prox-mapping is defined as $\text{Prox}_x(\xi) = \textbf{argmin}_{u \in X}\{V_w(u,x) + \langle \xi, u \rangle\}$, where w is 1-strongly convex.
3. **Mirror Descent**: do $x_{t+1} = \text{Prox}_{x_t}(\gamma_t g(x_t)) = \textbf{argmin}_{x \in X}\{w(x) + \langle \gamma_t g(x_t) - \nabla w(x_t), x \rangle\}$. Recall that the subgradient descent is equivalent to $x_{t+1} = \textbf{argmin}_{x \in X}\{\frac{1}{2}\|x - x_t\|^2 + \langle \gamma_t g(x_t), x \rangle\}$ and $V_w(x,y) = \frac{1}{2}\|x - y\|^2$ for $w(x) = \frac{1}{2}\|x\|^2$. Therefore, subgradient descent is a special case of mirror descent.

Properties:

1. **Three point identity**: (6.26) $V_w(x,z) = V_w(x,y) + V_w(y,z) - \langle \nabla w(z) - \nabla w(y), x - y \rangle$. Proof: by definition.
2. **Generalized Pythagorean Theorem for Bregman Divergence**: (6.23) if $x^* = \textbf{argmin}_{x \in C} V_w(x, x_0)$ for some convex set C, then for any $y \in C$, we have $V_w(y, x_0) \geq V_w(y, x^*) + V_w(x^*, x_0)$. Proof: use $\nabla V_w(x, x_0)^\top |_{x=x^*} (y - x^*) \geq 0 \Leftrightarrow (\nabla w(x^*) - \nabla w(x_0))^\top(y - x^*) \geq 0$ and the three point identity.
3. **Convergence of mirror descent**: (6.28) let f be convex, we have $\min_{1 \leq t \leq T} f(x_t) - f^* \leq \frac{1}{\sum_{t=1}^T \gamma_t}(V_w(x^*, x_1) + \frac{1}{2}\sum_{t=1}^T \gamma_t^2 \|g(x_t)\|^2)$. Proof: since $x_{t+1} = \textbf{argmin}_{x \in X}\{w(x) + \langle \gamma_t g(x_t) - \nabla w(x_t), x \rangle\}$, by optimality we have $\langle \nabla w(x_{t+1}) + \gamma_t g(x_t) - \nabla w(x_t), x - x_{t+1} \rangle \geq 0$. Therefore, we have $\langle \gamma_t g(x_t), x_{t+1} - x \rangle \leq \langle \nabla w(x_{t+1}) - \nabla w(x_t), x - x_{t+1} \rangle = V_w(x, x_t) - V_w(x, x_{t+1}) - V_w(x_{t+1}, x_t)$. Using $\langle \gamma_t g(x_t), x_t - x_{t+1} \rangle \leq \frac{\gamma_t^2}{2}\|g(x_t)\|^2 + \frac{1}{2}\|x_t - x_{t+1}\|^2$ and noticing $V_w(x_{t+1}, x_t) \geq \frac{1}{2}\|x_t - x_{t+1}\|^2$ by 1-strongly convexity, we have $\langle \gamma_t g(x_t), x_t - x^* \rangle \leq V_w(x^*, x_t) - V_w(x^*, x_{t+1}) + \frac{\gamma_t^2}{2}\|g(x_t)\|^2$. The rest is the same as the proof of (6.17).

7 Stochastic Optimization

The goal is to $\min_{x\in X} F(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$, or more generally, $\min_{x\in X} F(x) = \mathbb{E}_\xi(f(x,\xi))$. For large n, computing the full gradient is expensive. For unknown $P(\xi)$, the gradient is intractable.

Stochastic gradient quality for smooth functions: (12.12) let $F(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$, where f_i is L_i-smooth and convex, and F has a global minimum x^*. Let $L_{\max} = \max_i\{L_i\}$, then for any x we have $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x) - \nabla f_i(x^*)\|^2 \le 2L_{\max}(F(x) - F(x^*))$. Proof: define $g_i(x) = f_i(x) - f_i(x^*) - \nabla f_i(x^*)^\top(x-x^*)$. Thus, $g_i(x) \ge 0$, and is convex and L_i-smooth. By sufficient decrease, we have $0 \le g_i(x - \frac{1}{L_i}\nabla g_i(x)) \le g_i(x) - \frac{1}{2L_i}\|\nabla g_i(x)\|^2$. Thus, $g_i(x) \ge \frac{1}{2L_{\max}}\|\nabla g_i(x)\|^2$. Expanding this by definition, summing over i, and using $\nabla F(x^*) = 0$ yields the result. Note that the proof works for the general $F(x) = \mathbb{E}(x,\xi)$.

Stochastic gradient descent: do $x_{t+1} = \Pi_X(x_t - \gamma_t \nabla f(x_t,\xi_t))$. In the finite-sum problem, this is $x_{t+1} = \Pi_X(x_t - \gamma_t\nabla f_{i_t}(x_t))$, where i_t is sampled uniformly at random. Thus, the *gradient is unbiased*: $\mathbb{E}_{i_t}(\nabla f_{i_t}(x_t) \mid x_t) = \sum_i \frac{1}{n}\nabla f_i(x_t) = \nabla F(x_t)$, or $\mathbb{E}(\nabla f(x_t,\xi_t) \mid \xi_{[t-1]}) = \nabla F(x_t)$. The step size should diminish, i.e., $\gamma_t \to 0$, to ensure convergence, as the stochastic gradient does not necessarily equal to zero at the optima.

Analysis:

1. **Strongly convex functions**: (12.3) assume $F(x)$ is μ-strongly convex and $\mathbb{E}(\|\nabla f(x,\xi)\|^2) \le B^2$ for any $x \in X$. With $\gamma_t = \gamma/t$ for $\gamma > \frac{1}{2\mu}$, SGD satisfies $\mathbb{E}(\|x_t - x^*\|^2) \le \frac{C(\gamma)}{t}$, where $C(\gamma) = \max\{\frac{\gamma^2 B^2}{2\mu\gamma - 1}, \|x_1 - x^*\|^2\}$. Proof: by projection inequality, $\|x_{t+1} - x^*\|^2 \le \|x_t - \gamma_t\nabla f(x_t,\xi_t) - x^*\|^2 = \|x_t - x^*\|^2 - 2\gamma_t\langle\nabla f(x_t,\xi_t), x_t - x^*\rangle + \gamma_t^2\|\nabla f(x_t,\xi_t)\|^2$. Taking expectation, we have $\mathbb{E}(\|x_{t+1} - x^*\|^2) \le \mathbb{E}(\|x_t - x^*\|^2) - 2\gamma_t\mathbb{E}(\langle\nabla f(x_t,\xi_t), x_t - x^*\rangle) + \gamma_t^2 B^2$. Note that $\mathbb{E}(\langle\nabla f(x_t,\xi_t), x_t - x^*\rangle) = \mathbb{E}\Big[\mathbb{E}(\langle\nabla f(x_t,\xi_t), x_t - x^*\rangle \mid \xi_{[t-1]})\Big] = \mathbb{E}(\langle\nabla F(x_t), x_t - x^*\rangle)$. By strong convexity, $\langle\nabla F(x_t), x_t - x^*\rangle \ge \mu\|x_t - x^*\|^2$, thus $\mathbb{E}(\|x_{t+1} - x^*\|^2) \le (1 - 2\mu\gamma_t)\mathbb{E}(\|x_t - x^*\|^2) + \gamma_t^2 B^2$. The result follows by induction.
2. **Convex functions**: (12.4) let F be convex and $\mathbb{E}(\|\nabla f(x,\xi)\|^2) \le B^2$ for any $x \in X$. SGD satisfies $\mathbb{E}(F(\hat{x_T}) - F(x^*)) \le \frac{R^2 + B^2\sum_{t=1}^T \gamma_t^2}{2\sum_{t=1}^T \gamma_t}$ for $\hat{x_T} = \frac{\sum_{t=1}^T \gamma_t x_t}{\sum_{t=1}^T \gamma_t}$. Proof: use $\langle\nabla F(x_t), x_t - x^*\rangle \ge F(x_t) - F(x^*)$ in the proof of (12.3) gives $\gamma_t\mathbb{E}(F(x_t) - F(x^*)) \le \frac{1}{2}\mathbb{E}(\|x_t - x^*\|^2) - \frac{1}{2}\mathbb{E}(\|x_{t+1} - x^*\|^2) + \frac{1}{2}\gamma_t^2 B^2$. The result follows by recursion and convexity.
3. **Strongly convex and smooth functions, constant step size**: (12.5) assume $F(x)$ is both μ-strongly convex and L-smooth, and $\mathbb{E}(\|\nabla f(x,\xi)\|^2) \le \sigma^2 + c\|\nabla F(x)\|^2$. Then, with $\gamma_t = \gamma \le \frac{1}{Lc}$, $\mathbb{E}(F(x_t) - F(x^*)) \le \frac{\gamma L\sigma^2}{2\mu} + (1 - \gamma\mu)^{t-1}(F(x_1) - F(x^*))$.
4. **Non-convex but smooth function**: (12.8) assume $F(x) = \mathbb{E}(f(x,\xi))$ is L-smooth and $\mathbb{E}(\|\nabla f(x,\xi) - \nabla F(x)\|^2) \le \sigma^2$, then with $\gamma_t = \min\{1/L, \frac{\gamma}{\sigma\sqrt{T}}\}$, SGD achieves $\mathbb{E}(\|\nabla F(\hat{x}_T)\|^2) \le \frac{\sigma}{T}(2(F(x_1) - F(x^*))/\gamma + L\gamma)$, where $\hat{x}_T$ is selected uniformly at random from $\{x_1, \ldots, x_T\}$. Proof: by L-smoothness and $x_{t+1} = x_t - \gamma_t\nabla f(x_t,\xi_t)$, we have $\mathbb{E}(F(x_{t+1}) - F(x_t)) \le \mathbb{E}(\nabla F(x_t)^\top(x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|^2) \le \mathbb{E}(-\gamma_t\nabla F(x_t)^\top\nabla f(x_t,\xi_t) + \frac{L\gamma_t^2}{2}\|\nabla f(x_t,\xi_t)\|)$. Using $\mathbb{E}(\nabla f(x_t,\xi_t) \mid x_t) = \nabla F(x_t)$ and $\mathbb{E}(\|\nabla f(x_t,\xi_t)\|^2 \mid x_t) \le \sigma^2 + \|\nabla F(x_t)\|^2$, this implies $\mathbb{E}(F(x_{t+1}) - F(x_t)) \le -\frac{\gamma_t}{2}\mathbb{E}(\|\nabla F(x_t)\|^2) + \frac{L\sigma^2\gamma_t^2}{2}$ since $\gamma_t \le 1/L$. The result follows by induction.

Variants of SGD

1. **AdaGrad**: do $v_t = v_{t-1} + \nabla f(x_t,\xi_t)^{\odot 2}$ and $x_{t+1} = x_t - \frac{\gamma_0}{\epsilon + \sqrt{v_t}} \odot \nabla f(x_t,\xi_t)$, where $\odot$ means element-wise operation. Idea: adjust learning rate for different coordinate, use smaller step-size for mostly updates coordinates. Problem: learning rate is adjusted too aggressively and becomes too small at a later stage.
2. **RMSProp**: do $v_t = \beta v_{t-1} + (1-\beta)\nabla f(x_t,\xi_t)^{\odot 2}$ and $x_{t+1} = x_t - \frac{\gamma_0}{\epsilon + \sqrt{v_t}} \odot \nabla f(x_t,\xi_t)$. Idea: use a moving aqerage as the discount factor so that the discount factor does not grow too fast. β is chosen close to 1.
3. **Adam**: do $v_t = \beta_2 v_{t-1} + (1-\beta_2)\nabla f(x_t,\xi_t)^{\odot 2}$, $m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla f(x_t,\xi_t)$ and $x_{t+1} = x_t - \frac{\gamma_0}{\epsilon + \sqrt{v_t/(1-\beta_2^t)}} \odot \frac{m_t}{1-\beta_1^t}$. Idea: combining momentum with learning rate adjustment. β_1 and β_2 are chosen close to 1.

Variance Reduction Technique: the convergence guarantee of SGD with constant step size depends on the variance of the gradient. Reducing its variance makes it converge to a nearer point to the optimal.

1. **Mini-batch**: use a mini-batch to estimate gradient, i.e., $\nabla f(x_t,\xi_t) = \frac{1}{b}\sum_{i=1}^b \nabla f(x_t,\xi_{t,i})$.
2. **Momentum**: do $x_{t+1} = x_t - \gamma_t\hat{m}_t$, where $\hat{m}_t = c\sum_{\tau=1}^t \alpha^{t-\tau}\nabla f_{i_\tau}(x_\tau)$ is the weighted average of the past stochastic gradients.
3. **Control variate**: assume we want to estimate $\theta = \mathbb{E}X$, and we know a random variable Y that is highly correlated with X and $\mathbb{E}Y$ can be computed easily. Then $\hat{\theta}_\alpha := \alpha(X - Y) + \mathbb{E}Y$ has smaller bias and larger variance when α increases from 0 to 1. When $\alpha = 0$ it has zero variance and when $\alpha = 1$ it has zero bias.

Stochastic variance-reduced algorithms:

	SVRG	SAG/SAGA
memory cost	$O(d)$	$O(nd)$
epoch-based	yes	no
# gradients per step	at least 2	1
parameters	stepsize & epoch length	stepsize
unbiasedness	yes	yes/no
total complexity	$O\left((n+\kappa_{\max})\log\frac{1}{\epsilon}\right)$	$O\left((n+\kappa_{\max})\log\frac{1}{\epsilon}\right)$

1. **Stochastic average gradient (SAG)**: use average of the past gradient as an estimate of the full gradient: $g_t = \frac{1}{n}\sum_{i=1}^n v_i^t$, where $v_i^t = \nabla f_{i_t}(x_t)$ for $i = i_t$ and $v_i^t = v_i^{t-1}$ otherwise. Equivalently, $g_t = g_{t-1} + \frac{1}{n}(\nabla f_{i_t}(x_t) - v_{i_t}^{t-1})$. This means SAG is as cheap as SGD. It has convergence rate linear in $\log 1/\epsilon$ for strongly convex and smooth functions.
2. **SAGA**: use $g_t = \nabla f_{i_t}(x_t) - v_{i_t}^{t-1} + \frac{1}{n}\sum_{i=1}^n v_i^{t-1}$ to make it unbiased. It has the same rate as SAG.
3. **Stochastic variance-reduced gradient (SVRG)**: use a fixed reference point to estimate the gradient: $g_t = \nabla f_{i_t}(x_t) - \nabla f_{i_t}(\tilde{x}) + \nabla F(\tilde{x})$ and $x_{t+1} = x_t - \eta g_t$, where the reference point $\tilde{x}_t$ is updated only once a while. Idea: $\mathbb{E}(\|g_t - \nabla F(x_t)\|^2) \le \mathbb{E}(\|\nabla f_{i_t}(x_t) - \nabla f_{i_t}(\tilde{x})\|^2) \le L_{\max}^2\|x_t - \tilde{x}\|^2$, so the variance is bounded by how close x_t is to $\tilde{x}$. Typical choice of $\tilde{x}$: each epoch s consists of m update, then use $\tilde{x}^s = \frac{1}{m}\sum_i x_t$ as the reference point of next epoch.

Convergence of SVRG: (12.11) assume $f_i(x)$ is convex and L-smooth and $F(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ is μ-strongly convex. Let $x^* = \mathbf{argmin}_x F(x)$. For large m and $\eta < \frac{1}{2L}$, and $\rho := \frac{1}{\mu\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta} < 1$, we have $\mathbb{E}(F(\tilde{x}^s) - F(x^*)) \le \rho^s(F(\tilde{x}^0) - F(x^*))$. Proof: by $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, we have $\mathbb{E}(\|g_t\|^2) \le 2\mathbb{E}(\|\nabla f_{i_t}(x_t) - \nabla f_{i_t}(x^*)\|^2) + 2\mathbb{E}(\|\nabla f_{i_t}(\tilde{x}) - \nabla f_{i_t}(x^*) - \nabla F(\tilde{x}^*)\|^2)$. Notice $\mathbb{E}(\|\nabla f_{i_t}(\tilde{x}) - \nabla f_{i_t}(x^*) - \nabla F(\tilde{x}^*)\|^2) = \mathbb{E}(\|\nabla f_{i_t}(\tilde{x}) - \nabla f_{i_t}(x^*)\|^2)$ and use lemma 12.12, we have $\mathbb{E}(\|g_t\|^2) \le 4L(F(x_t) - F(x^*) + F(\tilde{x}) - F(x^*))$. Therefore, $\mathbb{E}(\|x_{t+1} - x^*\|^2) = \|x_t - x^*\|^2 - 2\eta(x_t - x^*)^\top\mathbb{E}(g_t) + \eta^2\mathbb{E}(\|g_t\|^2) \le \|x_t - x^*\|^2 - 2\eta(1 - 2L\eta)(F(x_t) - F(x^*)) + 4L\eta^2(F(\tilde{x}) - F(x^*))$. By convexity, we have $-\sum_{i=1}^m \mathbb{E}(F(x_t - F(x^*))) \le -m\mathbb{E}(F(\tilde{x}) - F(x^*))$. In addition, by strong convexity, we have $\mathbb{E}(\|\tilde{x} - x^*\|^2) \le \frac{2}{\mu}\mathbb{E}(F(\tilde{x}) - F(x^*))$. Combining these three gives the desired result. Set $\eta = O(1/L)$ and $m = O(L/\mu)$ makes $\rho \in (0, 0.5)$. Thus, the convergence is linear in $O(\log(1/\epsilon))$.

Other methods:

Parameters	SPIDER	SARAH	STORM
T	$\mathcal{O}(\varepsilon^{-2})$	$\mathcal{O}(\varepsilon^{-3})$	$\mathcal{O}(\varepsilon^{-3})$
T/Q	$\mathcal{O}(\varepsilon^{-1})$	$\mathcal{O}(\varepsilon^{-1})$	1
D	$\mathcal{O}(\varepsilon^{-2})$	$\mathcal{O}(\varepsilon^{-2})$	$\mathcal{O}(1)$
S	$\mathcal{O}(\varepsilon^{-1})$	$\mathcal{O}(1)$	$\mathcal{O}(1)$
η (or η_t)	0	0	$\mathcal{O}(t^{-2/3})$
α (or α_t)	$\mathcal{O}(1)$	$\mathcal{O}(\varepsilon)$	$\mathcal{O}(t^{-1/3})$
Complexity	$\mathcal{O}(\varepsilon^{-3})$	$\mathcal{O}(\varepsilon^{-3})$	$\tilde{\mathcal{O}}(\varepsilon^{-3})$

8 Nonconvex Functions

Smooth functions (no longer convex): f is called L-smooth if $f(y) \le f(x) + \nabla f(x)^\top(y-x) + \frac{L}{2}\|y-x\|^2$. (9.1) If $\|\nabla^2 f(x)\| \le L$ for any x, then f is smooth. For convex f and *open* $\mathbf{dom}(f)$, the reverse is also true.

Gradient Descent for smooth functions: (9.2) let f has a global minimum x^* and is L-smooth. With $\gamma = 1/L$, GD yields $\frac{1}{T}\sum_t\|\nabla f(x_t)\|^2 \le \frac{2L}{T}(f(x_0) - f(x^*))$. Proof: by sufficient decrease, we have $\|\nabla f(x_t)\|^2 \le 2L(f(x_t) - f(x_{t+1}))$. Sum together gives the

result.
GD with $\gamma = 1/L$ cannot overshoot: (9.3) if x is not a critical point, and f is L-smooth over the line connecting x and $x' = x - \gamma\nabla f(x)$. Then with $\gamma = 1/L' < 1/L$, x' is not a critical point. This means there is no critical point between x and x', so no overshooting.
Deciding whether a critical point is a local minimum is coNP-complete.

9 Frank-Wolfe Algorithm

This is to solve constrained optimization, similar to PGD.
Definitions:
1. **Linear minimization oracle (LMO)**: $\mathrm{LMO}_X(g) = \mathbf{argmin}_{z\in X}\, g^\top z$.
2. **Frank-Wolfe algorithm**: do $s = \mathrm{LMO}_X(\nabla f(x_t))$ and $x_{t+1} = (1-\gamma_t)x_t + \gamma_t s$. Benefits: (1) the iterates are always feasible for convex domain; (2) the algorithm is projection-free if we can solve LMO; (3) the iterates have a sparse representation, defined by the combination of LMO in the previous steps; (4) it is affine-invariant, meaning that affine equivalent problems can be solved at the same cost, i.e., if $g(x) = f(Ax+b)$ and $\mathbf{dom}(g) = A^{-1}(\mathbf{dom}(f) - b)$, then minimizing g is the same to minimizing f.
3. **Curvature constant**: the curvature constant of the constrained optimization is defined to be $C_{f,X} = \sup_{y=(1-\gamma)x+\gamma s, \gamma\in(0,1]} \frac{1}{\gamma^2}(f(y) - f(x) - \nabla f(x)^\top (y-x))$.

LASSO (L_1 domain): minimize $\|Ax-b\|^2$ subject to $\|x\|_1 \le 1$. Thus, $\mathrm{LMO}_X(g) = \mathbf{argmin}_{z=\pm e_i, i\in[n]}\, g^\top z = -\mathbf{sgn}(g_i)e_i$, where $i = \mathbf{argmax}|g_i|$. Therefore, we can compute LMO in $O(\log(d))$.

Examples	$\mathcal{A}$	$\lvert\mathcal{A}\rvert$	dim.	LMO_X (g)
L1-ball	$\{\pm\mathbf{e}_i\}$	$2d$	d	$\pm\mathbf{e}_i$ with $\mathrm{argmax}_i \lvert g_i\rvert$
Simplex	$\{\mathbf{e}_i\}$	d	d	$\mathbf{e}_i$ with $\mathrm{argmin}_i\, g_i$
Spectahedron	$\{\mathbf{x}\mathbf{x}^\top, \lVert\mathbf{x}\rVert = 1\}$	∞	d^2	$\mathrm{argmin}_{\lVert\mathbf{x}\rVert=1}\, \mathbf{x}^\top G\mathbf{x}$
Norms	$\{\mathbf{x}, \lVert\mathbf{x}\rVert \le 1\}$	∞	d	$\mathrm{argmin}_{\mathbf{s}, \lVert\mathbf{s}\rVert\le 1} \langle \mathbf{s}, \mathbf{g}\rangle$
Nuclear norm	$\{Y, \lVert Y\rVert_* \le 1\}$	∞	d^2	..
Wavelets	..	∞	∞	..

Duality gap as a certificate for optimization quality: We define the duality gap at x to be $g(x) = \nabla f(x)^\top (x-s)$, where $s = \mathrm{LMO}_X(\nabla f(x))$. (8.2) For convex f, $g(x) \ge f(x) - f(x^*)$. Proof: $g(x) = \nabla f(x)^\top x - \min_{z\in X} \nabla f(x)^\top z \ge \nabla f(x)^\top x - \nabla f(x)^\top x^* \ge f(x) - f(x^*)$.
Analysis:
1. **Decrease property**: (8.4) for $\gamma_t \in [0,1]$, we have $f(x_{t+1}) \le f(x_t) - \gamma_t g(x_t) + \gamma_t^2 \frac{L}{2}\|s - x_t\|^2$, where $g(x_t)$ is the duality gap. Proof: by smoothness, $f(x_{t+1}) \le f(x_t) + \nabla f(x_t)^\top \gamma_t (s - x_t) + \gamma_t^2 \frac{L}{2}\|s-x_t\|^2 = f(x_t) - \gamma_t g(x_t) + \gamma_t^2\frac{L}{2}\|s - x_t\|^2$.
2. **Affine-invariant decrease**: we have $f(x_{t+1}) \le f(x_t) - \gamma_t g(x_t) + \gamma_t^2 C_{f,X}$. Proof: plug in $x = x_t$, $y = (1-\gamma_t)x_t + \gamma_t s$ into the definition of curvature constant.
3. $\gamma_t = \frac{2}{t+2}$: (8.3) assume f is convex and L-smooth, and X is convex, closed and bounded, then $f(x_T) - f(x^*) \le \frac{2LR^2}{T+1}$, where $R = \max_{x,y\in X}\|x-y\|$. Proof: define $h(x) = f(x) - f(x^*)$, thus $h(x_{t+1}) \le h(x_t) - \gamma_t h(x_t) + \gamma_t^2\frac{L}{2}\|s - x_t\|^2 \le (1-\gamma_t)h(x_t) + \gamma_t^2 C$, where $C = \frac{L}{2}R^2$. By induction, we have $h(x_t) \le \frac{4C}{t+1}$. (8.5) The same holds for $C_{f,X}$ instead of C.
4. $\gamma_t' = \mathbf{argmin}_{\gamma\in[0,1]} f((1-\gamma)x_t + \gamma s)$: the same result holds as now we have $h(x_{t+1}) \le h((1-\gamma_t)x_t + \gamma_t s) \le (1-\gamma_t)h(x_t) + \gamma_t^2 C$.
5. $\gamma_t' = \min(\frac{g(x_t)}{L\|s-x_t\|^2}, 1)$: the same result holds as now we have $h(x_{t+1}) \le h(x_t) - \gamma_t g(x_t) + \gamma_t^2 C$ as well, since γ_t' minimizes this upper bound.
6. **Convergence of duality gap**: (8.7) under the same setting, $g(x_t) \le O(\frac{C_{f,X}}{T+1})$.

10 Newton's Method

Find zeros: $x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}$. This is to solve the first-order approximation: $f(x_t) + f'(x_t)(x - x_t) = 0$.

Find minimum: $x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)}$. This is to search zeros of f'. Generally, $x_{t+1} = x_t - \nabla^2 f(x_t)^{-1}\nabla f(x_t)$. (10.3) $x_{t+1} = \mathbf{argmin}\, f(x_t) + \nabla f(x_t)^\top (x - x_t) + \frac{1}{2}(x-x_t)^\top \nabla^2 f(x_t)(x-x_t)$.

Analysis:
1. **Nondegenerate quadratic function**: (10.1) for $f(x) = \frac{1}{2}x^\top M x - q^\top x + c$ with invertible symmetric M, Newton's method yields $x_1 = x^* = M^{-1}q$ for any x_0.
2. **Affine invariance**: (10.2) let f be twice differentiable, A invertible and $g(y) = Ay + b$. Define $N_h(x) = x - \nabla^2 h(x)^{-1}\nabla h(x)$, then $N_{f\circ g} = g^{-1}\circ N_f \circ g$.

$$\begin{tikzcd} \mathbf{x}_t \arrow[r, "N_f"] & \mathbf{x}_{t+1} \arrow[d, "g^{-1}"] \\ \mathbf{y}_t \arrow[u, "g"] \arrow[r, "N_{f\circ g}"] & \mathbf{y}_{t+1} \end{tikzcd}$$

3. **Bounded inverse Hessian and Lipschitz continuous Hessian**: (10.4) assume f is twice continuously differentiable, $\|\nabla^2 f(x)^{-1}\| \le 1/\mu$ and $\|\nabla^2 f(x) - \nabla^2 f(y)\| \le B\|x-y\|$ for any $x, y \in X$. Then we have $\|x_{t+1} - x^*\| \le \frac{B}{2\mu}\|x_t - x^*\|^2$ if a critical point x^* exists in X. Proof: Define $H(x) = \nabla^2 f(x)$. $x_{t+1} - x^* = x_t - x^* + H(x_t)^{-1}(\nabla f(x^*) - \nabla f(x_t)) = x_t - x^* + H(x_t)^{-1}\int_0^1 H(x_t + u(x^* - x_t))(x^* - x_t)du = H(x_t)^{-1}\int_0^1 (H(x_t + u(x^*-x_t)) - H(x_t))(x^* - x_t)du$. Therefore, we have $\|x_{t+1} - x^*\| \le \|H(x_t)^{-1}\|\|x^* - x_t\|\int_0^1 \|H(x_t + u(x^* - x_t)) - H(x_t)\|du \le \frac{1}{\mu}\|x_t - x^*\|^2 B\int_0^1 u\,du = \frac{B}{2\mu}\|x_t - x^*\|^2$.
4. **Fast convergence if near**: (10.5) under the assumption of (10.4), if $\|x_0 - x^*\| \le \mu/B$, then $\|x_T - x^*\| \le \frac{\mu}{B}(\frac{1}{2})^{2^T - 1}$. Proof: induction on (10.4).
5. **Global convergence for strongly convex and smooth functions**: with $\gamma = \mu/L$, (scaled) Newton's method $x_{t+1} = x_t - \gamma\nabla^2 f(x_t)^{-1}\nabla f(x_t)$ satisfies $f(x_t) - f^* \le (1 - \frac{\mu^2}{L^2})^t (f(x_0) - f^*)$. Note that the constant is worse than GD. Proof: expand $f(x_{t+1}) - f(x_t)$ by smoothness, then use $\frac{1}{L} \le \|H_t^{-1}\| \le \frac{1}{\mu}$ and $\|\nabla f(x_t)\|^2 \ge 2\mu(f(x_t) - f^*)$.

11 Quasi-Newton Methods

Computing and inverting the Hessian in Newton's method is costly. Quasi-Newton methods want to avoid this.
Secant condition: use approximation $f'(x_t) \approx \frac{f(x_t) - f(x_{t-1})}{x_t - x_{t-1}}$ to find zeros. Thus, the second method updates as: $x_{t+1} = x_t - f'(x_t)\frac{x_t - x_{t-1}}{f'(x_t) - f'(x_{t-1})}$. To generalize to higher dimensions, we want to find H_t such that $\nabla f(x_t) - \nabla f(x_{t-1}) = H_t(x_t - x_{t-1})$, and do $x_{t+1} = x_t - H_t^{-1}\nabla f(x_t)$.
Quasi-Newton method: if the method satisfies secant condition, we say it is a quasi-Newton method. In the multidimensional case, H_t is not unique. In particular, Newton's method is a quasi-Newton method iff f is a non-degenerate quadratic function. For efficiency, quasi-Newton methods typically deal with the inverse directly.
Greenstadt's approach: update $H_t^{-1} = H_{t-1}^{-1} + E_t$ for some symmetric E_t with small $\|E_t\|_F^2 = \sum_{i,j} e_{i,j}^2$. To introduce more flexibility, we minimize $\|AE_tA^\top\|_F^2$ for some fixed invertible matrix A. The program is: minimize $\frac{1}{2}\|AE_tA^\top\|_F^2$ such that $E_t^\top = E_t$ and $E_t(\nabla f(x_t) - \nabla f(x_{t-1})) = x_t - x_{t-1} - H_{t-1}^{-1}(\nabla f(x_t) - \nabla f(x_{t-1}))$, which is the secant condition. Note that this is a convex program, so we solve by Lagrange method. For simplicity, we write the program as: minimize $\frac{1}{2}\|AEA^\top\|_F^2$ such that $Ey = r$ and $E^\top - E = 0$. Define $f(E) = \frac{1}{2}\|AEA^\top\|_F^2$, thus $\nabla f(E) = A^\top AEA^\top A$. Note that the constraints are in fact linear over $e_{i,j}$. (11.3) Define $M = (A^\top A)^{-1}$ which is positive definite, then a solution E^* is optimal iff $E^* = M(\lambda y^\top + \Gamma^\top - \Gamma)M$, where $\lambda \in \mathbb{R}^{d\times 1}$ and $\Gamma \in \mathbb{R}^{d\times d}$. Solving the linear system gives $E^* = \frac{1}{y^\top My}(ry^\top M + Myr^\top - \frac{y^\top r}{y^\top My}Myy^\top M)$. This is called the Greenstadt method with parameter M.
BFGS: BFGS (named after four people) is the Greenstadt method with parameter H_t^{-1}. Note that H_t^{-1} is not yet known in the computation of E_t, but we have $H_t^{-1}y = x_t - x_{t-1}$ and in the formula of E^*, M appears in the form of My. This allows us to compute E^* without knowing the value of M. BFGS ensures that if f is not flat between x_{t-1} and x_t, and H_{t-1} is positive definite, then H_t is also positive definite. This reduces per iteration cost to $O(d^2)$.
L-BFGS: Define $\sigma = x_t - x_{t-1}$, $y = \nabla f(x_t) - \nabla f(x_{t-1})$, $H = H_{t-1}^{-1}$ and $H' = H_t^{-1}$, then BFGS can be written as $H' = (I - \frac{\sigma y^\top}{y^\top\sigma})H(I - \frac{y\sigma^\top}{y^\top\sigma}) + \frac{\sigma\sigma^\top}{y^\top\sigma}$. L-BFGS relies on the efficient computation of $H'g'$ given that Hg can be computed efficiently for any g and g'. This is because $H'g' = (I - \frac{\sigma y^\top}{y^\top\sigma})\left[H(I - \frac{y\sigma^\top}{y^\top\sigma})g'\right] + \sigma\frac{\sigma^\top g'}{y^\top\sigma}$ and note $(I - \frac{\sigma y^\top}{y^\top\sigma})s = s - \sigma\frac{y^\top s}{y^\top\sigma}$, thus can be computed in $O(d)$ with one call of Hg. Recur-

sively, we get $O(td)$ time complexity, which is not helpful. Instead, L-BFGS only does recursion for m times, and use H_0 instead of H_{t-m} (which is unknown to us as we do not explicitly compute this) as the result of m-th recursion. Note that we start close, so H_0 is close to H_{t-m}. In practice, we use a better choice than H_0. The final complexity is $O(md)$ per iteration.

12 Modern Second-order methods and nonconvex optimization

12.1 Cubic Regularization

Motivation: for L-Lipschitz Hessian as in the assumption of Newton's method, we have $f(x) \le f(x_t) + \nabla f(x_t)^\top (x - x_t) + \frac{1}{2}(x - x_t)^\top \nabla^2 f(x_t)(x - x_t) + \frac{L}{6}\|x - x_t\|^3$. Similar to GD which minimizes the upper bound given by smoothness (Lipschitz gradient), cubic regularization is to minimize the upper bound given by Lipschitz Hessian. Algorithm: $x_{t+1} = \mathbf{argmin}_x\, f(x_t) + \nabla f(x_t)^\top (x - x_t) + \frac{1}{2}(x - x_t)^\top \nabla^2 f(x_t)(x - x_t) + \frac{L}{6}\|x - x_t\|^3$. This can be reduced to a convex problem.

Convergence rate: $\min_{i \le t} \|\nabla f(x_i)\| = O(t^{-2/3})$. If convex, we have $f(x_t) - f^* = O(t^{-2})$.

GD, AGD	Cubic Regularization
▶ Use only gradient	▶ Use both gradient and Hessian
▶ Cheap $O(d)$ per-iteration cost	▶ Moderate $O(d^2)$ iteration cost
▶ Global convergence (requires gradient Lipschitzness)	▶ Faster global convergence (requires Hessian Lipschitzness)
▶ Convergence to first-order stationary point	▶ Convergence to second-order stationary point

12.2 Nonconvex Optimization

A nonconvex optimization may have exponentially many local minima, and determining whether a critical point is a local minimum is co-NP complete.

Classification of stationary points: (1) If $\nabla^2 F(x) \succ 0$, then x is a local minimum; (2) If $\nabla^2 F(x) \prec 0$, then x is a local maximum; (3) If $\nabla^2 F(x)$ has both positive and negative eigenvalues, then x is a (strict) saddle point; (4) Otherwise, it remains inconclusive.

Analysis:

1. **SGD converges to a stationary point**: see (12.8).
2. **SGD with random initialization**: with random initialization, GD converges to a local minimum almost surely.
3. **Noisy SGD**: with extra noise added to the SGD update, if f satisfies strict saddle property and has Lipschitz Hessian, then noisy SGD converges to a second-order stationary point.
4. **Benign landscapes**: if the function satisfies PL condition, or all local minimum is global minimum, then GD (in the latter case requires random initialization) converges to the global minimum.

13 Smoothing Techniques

13.1 Convex Conjugate

Convex conjugate (Legendre-Fenchel): for any f, its convex conjugate is given as $f^*(y) = \sup_{x \in \mathbf{dom}(f)}\{x^\top y - f(x)\}$.

Properties:

1. **Fenchel's inequality**: $f(x) + f^*(y) \ge x^\top y$ for any x, y. Proof by definition.
2. **Conjugate of a conjugate**: (7.2) If f is convex, lower semi-continuous and proper, then $(f^*)^* = f$. A proper convex function means $f(x) > -\infty$.
3. **Strong convexity implies nice conjugate**: (7.3) If f is μ-strongly convex, then f^* is continuously differentiable and $\frac{1}{\mu}$ smooth.
4. **Conjugate of function sum**: If f and g is lower semi-continuous and convex, then $(f+g)^*(x) = \inf_y\{f^*(y) + g^*(x-y)\}$.

13.2 Nesterov's Smoothing

Use $f_\mu(x) = \max_{y \in \mathbf{dom}(f^*)}\{x^\top y - f^*(y) - \mu d(y)\}$ as the surrogate function to minimize. $d(y)$ is a 1-strongly convex and nonnegative everywhere function, called the proximity function. Typical choices include $d(y) = \frac{1}{2}\|y - y_0\|^2$ and $d(y) = \frac{1}{2}\sum w_i (y_i - y_{0,i})^2$ for $w_i \ge 1$. This makes $f_\mu(x)$ to be $\frac{1}{\mu}$-smooth, as $f^*(y) + \mu d(y)$ is μ-strongly convex.

Approximation Error: for convex f with bounded $\mathbf{dom}(f^*)$, we have $f(x) - \mu D^2 \le f_\mu(x) \le f(x)$, where $D^2 = \max_{y \in \mathbf{dom}(f^*)} d(y)$. The tradeoff between approximation error and optimization error: $f(x) - f^* \le [f(x) - f_\mu(x)] + [f_\mu(x) - \min_x f_\mu(x)]$. With AGD, we get $f(x_t) - f^* = O(\mu D^2 + \frac{R^2}{\mu t^2})$. To achieve approximation error ϵ, we need $\mu = O(\frac{\epsilon}{D^2})$. Therefore, $T_\epsilon = O(\frac{RD}{\epsilon})$.

13.3 Moreau-Yosida Smoothing

Use $f_\mu(x) = \min_y\{f(y) + \frac{1}{2\mu}\|x - y\|^2\}$. This is actually the special case of Nesterov's smoothing with $d(y) = \frac{1}{2}\|y\|^2$. Proof: $f_\mu^{Nes}(x) = \max_y\{x^\top y - f^*(y) - \frac{\mu}{2}\|y\|^2\} = (f^* + \frac{\mu}{2}\|\cdot\|^2)^*(x) = \inf_y\{f(y) + \frac{1}{2\mu}\|x - y\|^2\}$.

Properties: (1) $f_\mu(x)$ is $\frac{1}{\mu}$-smooth; (2) $\min_x f(x) = \min_x f_\mu(x)$. (3) GD reduces to proximal minimization: define $\mathbf{prox}_{\mu \cdot f} = \mathbf{argmin}_y\{f(y) + \frac{1}{2\mu}\|x - y\|^2\}$, then $x_{t+1} = x_t - \mu \nabla f_\mu(x_t) \Leftrightarrow x_{t+1} = \mathbf{prox}_{\mu \cdot f}(x_t)$.

Proximal operator: the proximal operator of convex function f at x is defined as $\mathbf{prox}_f(x) = \mathbf{argmin}_y\{f(y) + \frac{1}{2}\|x - y\|^2\}$. For continuous convex function f, $\mathbf{prox}_f(x)$ exists and is unique.

Properties:

1. **Fixed point**: If f is convex, then $x^* \in \mathbf{argmin}\, f(x) \Leftrightarrow x^* = \mathbf{prox}_f(x^*)$. Proof by definition.
2. **Non-expansive**: $\|\mathbf{prox}_f(x) - \mathbf{prox}_f(y)\| \le \|x - y\|$. Proof: Let $\mu_x = \mathbf{prox}_f(x)$ and $\mu_y = \mathbf{prox}_f(y)$. By optimality condition, $x - \mu_x \in \partial f(\mu_x)$ and $y - \mu_y \in \partial f(\mu_y)$. By the monotonicity of gradient, $(x - \mu_x - (y - \mu_y))^\top (\mu_x - \mu_y) \ge 0$. This means $\|\mu_x - \mu_y\|^2 \le (x - y)^\top (\mu_x - \mu_y) \le \|x - y\| \|\mu_x - \mu_y\|$.
3. **Moreau decomposition**: For any x, $x = \mathbf{prox}_f(x) + \mathbf{prox}_{f^*}(x)$. Proof: use $u_x \in \partial f^*(x - u_x)$.

Proximal Point Algorithm: $x_{t+1} = \mathbf{prox}_{\gamma_t \cdot f}(x_t)$. (7.14) If f is convex, then $f(x_T) - f^* \le \frac{\|x_0 - x^*\|^2}{2\sum_{t=0}^{T-1}\gamma_t}$. Proof: by definition, $f(x_{t+1}) + \frac{1}{2\gamma_t}\|x_{t+1} - x_t\|^2 \le f(x_t)$, which implies $f(x_{t+1}) - f(x_t) \le -\frac{1}{2\gamma_t}\|x_{t+1} - x_t\|^2$. By optimality condition, $0 \in \partial f(x_{t+1}) + \frac{1}{\gamma_t}(x_{t+1} - x_t)$, which implies $\frac{x_t - x_{t+1}}{\gamma_t} \in \partial f(x_{t+1})$. Therefore, by the definition of subgradient, $f(x_{t+1}) - f^* \le \frac{1}{\gamma_t}(x_t - x_{t+1})^\top (x_{t+1} - x^*) \le \frac{1}{\gamma_t}[(x_t - x^*)^\top (x_{t+1} - x^*) - \|x_{t+1} - x^*\|^2]$. Using $(x_t - x^*)^\top (x_{t+1} - x^*) \le \frac{1}{2}(\|x_t - x^*\|^2 + \|x_{t+1} - x^*\|^2)$, we get $f(x_{t+1}) - f^* \le \frac{1}{2\gamma_t}[\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2]$. The rest follows from non-increasing $f(x_t)$ and summing this inequality by t.

13.4 Lasry-Lions Smoothing

Use $f_{\mu,\delta}(x) = \max_y \min_z\{f(z) + \frac{1}{2\mu}\|z - y\|^2 - \frac{1}{2\delta}\|y - x\|^2\}$. This is double application of Moreau smoothing with function flipping. If f is 1-Lipschitz, then choose $\delta, \mu = O(\epsilon)$ guarantees ϵ approximation error, and $f_{\mu,\delta}$ is $O(1/\epsilon)$-smooth. This can be applied to nonconvex functions, but has computation inefficiency in this case.

13.5 Randomized Smoothing

Use $f_\mu(x) = \mathbb{E}_z[f(x + \mu z)]$, where z is an isotopic Gaussian or uniform random variable. Choosing $\mu = O(\epsilon)$ guarantees ϵ approximation error, and $f_\mu(x)$ is $O(\frac{\sqrt{d}}{\epsilon})$-smooth.

14 Min-Max Optimization

The min-max problem is defined as $\min_{x \in X} \max_{y \in Y} \phi(x, y)$.

Definitions:

1. **Saddle point and minimax point**: (x^*, y^*) is a saddle point if $\phi(x^*, y) \le \phi(x^*, y^*) \le \phi(x, y^*)$ for any x, y. (x^*, y^*) is a global minimax point if $\phi(x^*, y) \le \phi(x^*, y^*) \le \max_{y' \in Y} \phi(x, y')$ for any x, y. Game theory interpretation: saddle point means Nash equilibrium, where no play has the incentive to make unilateral changes; global minimax point is Stackelberg equilibrium, where it is the best response to the best response.
2. **Convex-concave function**: $\phi(x, y)$ is convex-concave if $\phi(x, y)$ is convex for every fixed y and $\phi(x, y)$ is concave for every fixed x.
3. **Strongly-convex-strongly-concave function**: $\phi(x, y)$ is strongly-convex-strongly-concave if $\phi(x, y)$ is μ_1-strongly convex for every fixed y and $\phi(x, y)$ is μ_2-strongly concave for every fixed x.
4. **Duality gap**: defined to be $\max_y \phi(\hat{x}, y) - \min_x \phi(x, \hat{y})$. When duality gap is smaller than or equal to ϵ, we say $(\hat{x}, \hat{y})$ is an ϵ-saddle point.

Properties:

1. **Characterization of saddle points**: Define $\overline{\phi}(x) = \max_y \phi(x, y)$ and $\underline{\phi}(y) = \min_x \phi(x, y)$. Then (x^*, y^*) is a saddle point iff $\max_y \min_x \phi(x, y) = \min_x \max_y \phi(x, y)$, $x^* \in \mathbf{argmin}_x \overline{\phi}(x)$ and $y^* \in \mathbf{argmax}_y \underline{\phi}(y)$. Proof: by definition.
2. **Minimax Theorem**: If X and Y are closed convex sets, one of them is bounded, and $\phi(x, y)$ is a continuous convex-concave function, then there exists a saddle

point on $X \times Y$ and $\max_y \min_x \phi(x, y) = \min \max \phi(x, y)$.

Gradient Descent Ascent: do $x_{t+1} = \Pi_X(x_t - \gamma \nabla_x \phi(x_t, y_t))$ and $y_{t+1} = \Pi_Y(y_t + \gamma \nabla_y \phi(x_t, y_t))$. This simultaneously updates x and y.