THINK BEFORE STARTING THE WRITING OF A PROOF. THINK OF ALL THE NECESSARY COMPONENTS FIRST. THERE IS ENOUGH TIME.

Definitions

- **Differentiable**: $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable if
$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + r(\boldsymbol{y} - \boldsymbol{x}),$$
where $\lim_{\boldsymbol{v}\to\boldsymbol{0}} \frac{|r(\boldsymbol{v})|}{\|\boldsymbol{v}\|} = 0$.
- **Spectral norm**: $\|A\|_2 = \sup_{\|\boldsymbol{x}\|=1} \|A\boldsymbol{x}\|$ (largest eigenvalue).
- **Positive semi-definite**: $\forall \boldsymbol{x} \in \mathbb{R}^d$: $\boldsymbol{x}^\top A \boldsymbol{x} \geq 0$.
- **Directional derivative**: If f is diff., $\langle \nabla f(\boldsymbol{x}), \boldsymbol{v}\rangle = \lim_{h\to 0} \frac{f(\boldsymbol{x}+h\boldsymbol{v}) - f(\boldsymbol{x})}{h}$.
- B-**Lipschitz**: $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}(f)$,
  [1] $\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \leq B\|\boldsymbol{x} - \boldsymbol{y}\|$.
  [2] If f differentiable, $\|\nabla f(\boldsymbol{x})\| \leq B$.
  [3] If f convex, $\|\boldsymbol{g}\| \leq B$, $\forall \boldsymbol{g} \in \partial f(\boldsymbol{x})$ (proof: subgrad def $\Rightarrow$ Cauchy-Schwarz).
- **Convex set**: $\forall \boldsymbol{x}, \boldsymbol{y} \in X, \lambda \in [0,1]$: $\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in X$.
- **Cone**: X is a cone if $\forall \boldsymbol{x} \in X, \lambda > 0$: $\lambda \boldsymbol{x} \in X$.
- **Convexity**: $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}(f)$ and $\forall \lambda \in [0,1]$,
  [1] $f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y})$.
  [2] $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle$.
  [3] $\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \geq 0$.
  [4] $\nabla^2 f(\boldsymbol{x})$ is positive semi-definite.
- **Convexity preservation**: Positive scaling, Sum, Max, and $f(A\boldsymbol{x} + \boldsymbol{b})$.
- L-**smoothness**: $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}(f)$,
  [1] $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$.
  [2] $g(\boldsymbol{x}) := \frac{L}{2}\|\boldsymbol{x}\|^2 - f(\boldsymbol{x})$ is convex.
  [3] $f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$ (canonical).
  [4] $\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \leq L\|\boldsymbol{x} - \boldsymbol{y}\|^2$.
  [5] $\|\nabla^2 f(\boldsymbol{x})\|_2 \leq L$.
  [6] If f is convex and L-smooth, then f is $1/L$-strongly convex:
  $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{1}{2L}\|\boldsymbol{x} - \boldsymbol{y}\|^2$.
  [7] Coordinate-wise: $f(\boldsymbol{x} + \lambda \boldsymbol{e}_i) \leq f(\boldsymbol{x}) + \lambda \nabla_i f(\boldsymbol{x}) + \frac{L_i}{2}\lambda^2, \forall \lambda \in \mathbb{R}$.
  Relations: $[5] \Leftrightarrow [1] \Rightarrow [2] \Leftrightarrow [3] \Leftrightarrow [4]$ (If convex, all $\Leftrightarrow$).
- **Smoothness preservation**: Pos. scaling scales, Sum sums. $f(A\boldsymbol{x} + \boldsymbol{b})$ has $L\|A\|_2^2$.
- μ-**strong convexity**: $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}(f)$,
  [1] $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$ (canonical).
  [2] $g(\boldsymbol{x}) := f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|^2$ is convex.
  [3] $\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \geq \mu\|\boldsymbol{x} - \boldsymbol{y}\|^2$ (proof: sum [1] for $(\boldsymbol{x}, \boldsymbol{y})$ and $(\boldsymbol{y}, \boldsymbol{x})$).
  [4] μ-SC $\Rightarrow$ PL inequality: $\frac{1}{2}\|\nabla f(\boldsymbol{x})\|^2 \geq \mu(f(\boldsymbol{x}) - f^\star)$.
- **Subgradient**: $\boldsymbol{g} \in \partial f(\boldsymbol{x}) \Leftrightarrow f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x}\rangle, \forall \boldsymbol{y} \in \mathrm{dom}(f)$.
- **Conjugate function**: $f^\star(\boldsymbol{y}) := \sup_{\boldsymbol{x}\in\mathrm{dom}(f)} \langle \boldsymbol{x}, \boldsymbol{y}\rangle - f(\boldsymbol{x})$.
- **Dual norm**: $\|\boldsymbol{y}\|_\star := \max_{\|\boldsymbol{x}\|\leq 1} \langle \boldsymbol{x}, \boldsymbol{y}\rangle$.

Lemmas

- $\frac{\mathrm{d}}{\mathrm{d}x}\frac{f(x)}{g(x)} = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$.
- **Cosine theorem**: All equivalent formulations,
  [1] $\|\boldsymbol{x} - \boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - 2\langle \boldsymbol{x}, \boldsymbol{y}\rangle$.
  [2] $\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \frac{1}{2}\left(\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - \|\boldsymbol{x} - \boldsymbol{y}\|^2\right)$.
  [3] $\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{z}\rangle = \frac{1}{2}\left(\|\boldsymbol{x} - \boldsymbol{y}\|^2 + \|\boldsymbol{x} - \boldsymbol{z}\|^2 - \|\boldsymbol{y} - \boldsymbol{z}\|^2\right)$.
- **Cauchy-Schwarz**:
  [1] $|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\|$.
  [2] $\left(\sum_{i=1}^n a_i b_i\right)^2 \leq \left(\sum_{i=1}^n a_i^2\right)\left(\sum_{i=1}^n b_i^2\right)$.
  [3] **Titu's lemma** ($b_i \geq 0$): $\frac{\left(\sum_{i=1}^n a_i\right)^2}{\sum_{i=1}^n b_i} \leq \sum_{i=1}^n \frac{a_i^2}{b_i}$ (proof: $a_i' = \frac{a_i}{\sqrt{b_i}}, b_i' = \sqrt{b_i}$).
- **Hölder's inequality** (special case): $|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \leq \|\boldsymbol{x}\|_1\|\boldsymbol{y}\|_\infty$.
- **Parallelogram law**: $2\|\boldsymbol{x}\|^2 + 2\|\boldsymbol{y}\|^2 = \|\boldsymbol{x} + \boldsymbol{y}\|^2 + \|\boldsymbol{x} - \boldsymbol{y}\|^2$.
- **Jensen's inequality** (φ convex, $a_i \geq 0$): $\varphi\left(\frac{\sum_{i=1}^m a_i \boldsymbol{x}_i}{\sum_{i=1}^m a_i}\right) \leq \frac{\sum_{i=1}^m a_i \varphi(\boldsymbol{x}_i)}{\sum_{i=1}^m a_i}$.
- **Fenchel's inequality**: $\langle \boldsymbol{x}, \boldsymbol{y}\rangle \leq f(\boldsymbol{x}) + f^\star(\boldsymbol{x}) \Rightarrow \langle \boldsymbol{x}, \boldsymbol{y}\rangle \leq \frac{1}{2}\left(\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|_\star^2\right)$.
- **Young's inequality** ($a, b \geq 0, \frac{1}{p} + \frac{1}{q} = 1$): $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$
  $\Rightarrow \|\boldsymbol{x}\|\|\boldsymbol{y}\| \leq \frac{1}{2}\left(\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2\right)$.
- $\frac{1}{\sqrt{d}}\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_1 \leq \sqrt{d}\|\boldsymbol{x}\|_2$.
- $\|A\boldsymbol{x}\| \leq \|A\|_2\|\boldsymbol{x}\|$.
- $\|A\|_2 \leq \|A\|_F$.
- **Mean-value theorem** (h cont. on $[a,b]$, diff. on (a,b)):
$$h'(c) = \frac{h(b) - h(a)}{b - a}, \quad \exists c \in (a,b).$$
- **Fund. theorem of calculus** (h diff. on $[a,b]$, h' cont. on $[a,b]$):
$$h(b) - h(a) = \int_a^b h'(t)\mathrm{d}t.$$
- $\left\|\int_0^1 \nabla h(t)\mathrm{d}t\right\| \leq \int_0^1 \|\nabla h(t)\|\mathrm{d}t$.
- $\int_0^1 c\mathrm{d}t = c, \quad \int_0^1 t\mathrm{d}t = \frac{1}{2}$.
- **Subgradient calculus**:
  [1] $h(\boldsymbol{x}) = \alpha f(\boldsymbol{x}) + \beta g(\boldsymbol{x}) \Rightarrow \partial h(\boldsymbol{x}) = \alpha \cdot \partial f(\boldsymbol{x}) + \beta \cdot \partial g(\boldsymbol{x})$.
  [2] $h(\boldsymbol{x}) = f(A\boldsymbol{x} + \boldsymbol{b}) \Rightarrow \partial h(\boldsymbol{x}) = A^\top \partial f(A\boldsymbol{x} + \boldsymbol{b})$.
  [3] $h(\boldsymbol{x}) = \max f_i(\boldsymbol{x}) \Rightarrow \partial h(\boldsymbol{x}) = \mathrm{conv}(\{\partial f_i(\boldsymbol{x}) \mid f_i(\boldsymbol{x}) = h(\boldsymbol{x})\})$.
- If f is differentiable at $\boldsymbol{x}$, then $\partial f(\boldsymbol{x}) \subseteq \{\nabla f(\boldsymbol{x}_t)\}$.
- If f is convex, then $\partial f(\boldsymbol{x}) \neq \emptyset$ for all in $\boldsymbol{x}$ in the relative interior.
- If $\mathrm{dom}(f)$ convex and $\partial f(\boldsymbol{x}) \neq \emptyset, \forall \boldsymbol{x} \in \mathrm{dom}(f)$, then f is convex.
- If f is strictly concave, the subgradient exists nowhere.
- For $p \geq 1$, $\frac{1}{p} + \frac{1}{q} = 1$, we have dual norms, $\|\cdot\|_{p,\star} = \|\cdot\|_q$.

Optimality lemmas (assume convexity)

The constrained and non-diff. cases are useful when update rule contains argmin.

- $\boldsymbol{x}^\star$ is a local minimum: $\exists \epsilon > 0$ such that $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{y}), \forall \boldsymbol{y} : \|\boldsymbol{x}^\star - \boldsymbol{y}\| \leq \epsilon$.
- $\nabla f(\boldsymbol{x}^\star) = \boldsymbol{0}$.
- Constrained: $\langle \nabla f(\boldsymbol{x}^\star), \boldsymbol{x} - \boldsymbol{x}^\star\rangle \geq 0, \forall \boldsymbol{x} \in X$.
- Non-differentiable: $\boldsymbol{0} \in \partial f(\boldsymbol{x}^\star)$.

Common tricks

- **Rearrange the update rule** for an equality. E.g., $\nabla f(\boldsymbol{x}_t) = \frac{\boldsymbol{x}_t - \boldsymbol{x}_{t+1}}{\gamma_t}$.
- Define $h(t) := f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))$, where $h'(t) = \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))^\top(\boldsymbol{y} - \boldsymbol{x})$ and use with FTOC: $f(\boldsymbol{y}) - f(\boldsymbol{x}) = \int_0^1 \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))^\top(\boldsymbol{y} - \boldsymbol{x})\mathrm{d}t$.
  Or, mean-value theorem: $\exists c \in (0,1) : \nabla f(\boldsymbol{x} + c(\boldsymbol{y} - \boldsymbol{x}))^\top(\boldsymbol{y} - \boldsymbol{x}) = f(\boldsymbol{y}) - f(\boldsymbol{x})$.
- **Projection is non-expansive**: $\|\Pi_X(\boldsymbol{x}) - \Pi_X(\boldsymbol{y})\| \leq \|\boldsymbol{x} - \boldsymbol{y}\|$.
- $\min_{1\leq t\leq T} f(\boldsymbol{x}_t) - f^\star \leq \frac{\sum_{t=1}^T \gamma_t(f(\boldsymbol{x}_t) - f^\star)}{\sum_{t=1}^T \gamma_t}$.
- **Telescoping sum** inequality: $\sum_{t=1}^T \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 \leq \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2$.
- A monotone and bounded sequence has a limit.
- If a value α is unknown for an algorithm. Start with a lower bound (or just $\tilde{\alpha}_0 = 1$) and run the algorithm, doubling every time $\tilde{\alpha}_{t+1} = 2 \cdot \tilde{\alpha}_t$ it is incorrect. This does not increase complexity because, in the end, $\tilde{\alpha}_T \leq 2\alpha$ and all the previous values with their iterations are a constant factor, smaller than the final run.
- To find the optimal $\gamma^\star$ that minimizes bound $q(\gamma)$, use 1st-order opt: $q(\gamma^\star) \stackrel{!}{=} 0$.
- $\max\{a, b\} \leq a + b$ if $a, b \geq 0$.
- $\sum_{t=1}^T \frac{1}{\sqrt{t}} = \mathcal{O}(\sqrt{T}), \quad \sum_{t=1}^T \frac{1}{t} = \mathcal{O}(\log T)$.
- $\|\boldsymbol{x}\| = \|\boldsymbol{x} - \boldsymbol{y} + \boldsymbol{y}\| \leq \|\boldsymbol{x} - \boldsymbol{y}\| + \|\boldsymbol{y}\|$, $\|\boldsymbol{x} - \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$
  $\Rightarrow \|\boldsymbol{x} - \boldsymbol{y}\| - \|\boldsymbol{y}\| \leq \|\boldsymbol{x}\| \leq \|\boldsymbol{x} - \boldsymbol{y}\| + \|\boldsymbol{y}\|$.
- $1 - x \leq \exp(-x), \forall x \geq 0 \Rightarrow (1-x)^y \leq \exp(-xy), \forall x \geq 0, y \in \mathbb{R}$.

IMPORTANT TIPS TO KEEP IN MIND

- When showing convexity, make sure to show that the domain is a convex set.
- If f is convex and want to use the subgradient, state that it exists bc of convexity.
- If something is obviously false, still provide a counterexample.
- Keep in mind divisions by 0 when defining functions. For example, when dividing by norm. Then, the gradient is not defined $\Rightarrow$ Use subgradient.
- Structure of a proof:
  [1] State what needs to be shown exactly and mark by $(\star)$.
  [2] State the assumptions of the question and their implications (think about which implications are relevant to the proof).
  [3] Proof should follow easily: "Hence, $(\star)$ holds and the proof is concluded.".
- If need to show that something does not exist, use proof by contradiction.
- If γ_t is timestep-dependent, generally need to use induction.

Expectation and variance for SGD

- $\mathrm{Var}[\boldsymbol{X}] := \mathbb{E}\left[\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]\|^2\right] = \mathbb{E}\left[\|\boldsymbol{X}\|^2\right] - \|\mathbb{E}[\boldsymbol{X}]\|^2$.
$$\Rightarrow \quad \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t, \boldsymbol{\xi}_t)\|^2\right] = \|\nabla F(\boldsymbol{x}_t)\|^2 + \mathbb{E}\left[\|\nabla f(\boldsymbol{x}_t, \boldsymbol{\xi}_t) - \nabla F(\boldsymbol{x}_t)\|^2\right] \leq \|\nabla F(\boldsymbol{x}_t)\|^2 + \sigma^2.$$
- **Law of total expectation**: $\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_X[X \mid Y]]$.
- **Law of total variance**: $\mathrm{Var}[Y] = \mathbb{E}_X[\mathrm{Var}_Y[Y \mid X]] + \mathrm{Var}_Y[\mathbb{E}_X[Y \mid X]]$.
- $\mathrm{Var}[X - Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] - 2 \cdot \mathrm{Cov}(X, Y)$.
- $\mathrm{Var}[\alpha X] = \alpha^2 \mathrm{Var}[X]$, $\mathrm{Var}[X + \beta] = \mathrm{Var}[X]$.

Risk minimization

- Unknown distribution P. We only have access to samples $X_1, \dots, X_n \sim P$. We want to explain data source X through these samples by minimizing risk.
- **Expected risk**: $\ell(H) := \mathbb{E}_X[\ell(H, X)]$.
- **Empirical risk**: $\ell_n(H) := \frac{1}{n}\sum_{i=1}^n \ell(H, X_i)$.
- **Probably approximately correct (PAC)**: Let $\epsilon, \delta > 0$, $\tilde{H} \in \mathcal{H}$ is PAC if, with probability at least $1-\delta$, $\ell(\tilde{H}) \leq \inf_{H\in\mathcal{H}} \ell(H) + \epsilon$.
- **Weak law of large numbers (WLLM)**: Let $H \in \mathcal{H}$ be fixed. For any $\delta, \epsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that for $n \geq n_0$, $|\ell_n(H) - \ell(H)| \leq \epsilon$ with probability at least $1-\delta$.
- Assume that for any $\delta, \epsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that for $n \geq n_0$, $\sup_{H\in\mathcal{H}} |\ell_n(H) - \ell(H)| \leq \epsilon$ with probability at least $1-\delta$. (WLLM holds uniformly for all hypotheses.) Then, an approximate empirical risk minimizer $\tilde{H}_n$ ($\ell_n(\tilde{H}_n) \leq \inf_{H\in\mathcal{H}} \ell_n(H) + \epsilon$) is PAC for expected risk minimization, meaning $\ell(\tilde{H}_n) \leq \inf_{H\in\mathcal{H}} \ell(H) + 3\epsilon$ with probability at least $1-\delta$.

$\ell(\tilde{H}_n) \overset{\text{uniform WLLM}}{\leq} \ell_n(\tilde{H}_n) + \epsilon \overset{\text{emp. risk min.}}{\leq} \inf_{H\in\mathcal{H}} \ell_n(H) + 2\epsilon \overset{\text{uniform WLLM}}{\leq} \inf_{H\in\mathcal{H}} \ell(H) + 3\epsilon$. $\square$

- **Empirical risk minimization** ($\ell_n(H_n)$: empirical, training; $\ell(H_n)$: expected, validation): We want generalization and learning,
  - (Low $\ell_n(H_n)$, High $\ell(H_n)$): Overfitting (theory is too complex).
  - (High $\ell_n(H_n)$, High $\ell(H_n)$): Underfitting (theory is too simple).
  - (Low $\ell_n(H_n)$, Low $\ell(H_n)$): Learning.
  - ($\ell_n(H_n) \approx \ell(H_n)$): Generalization.
  - Regularization: Punish complex hypotheses.
  - W.h.p. we do not have high $\ell_n(H_n)$, low $\ell(H_n)$, because $\ell_n(H_n) \leq \inf_{H\in\mathcal{H}} \ell_n(H) + \epsilon \leq \ell_n(\tilde{H}) + \epsilon \leq \ell(\tilde{H}) + 2\epsilon \leq \ell(\tilde{H}_n) + 3\epsilon$.

Non-linear programming

- **Optimization problem**:

$$\begin{aligned} &\text{minimize} && f_0(\boldsymbol{x}) \\ &\text{subject to} && f_i(\boldsymbol{x}) \leq 0, \quad i \in [m] \\ & && h_j(\boldsymbol{x}) = 0, \quad j \in [p]. \end{aligned}$$

- **Problem domain**: $X = \left(\bigcap_{i=0}^m \operatorname{dom}(f_i)\right) \cap \left(\bigcap_{j=1}^p \operatorname{dom}(h_j)\right)$.
- **Convex program**: All f_i are convex and all h_j are affine with domain $\mathbb{R}^d$.
- **Lagrangian**: $L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) := f_0(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i f_i(\boldsymbol{x}) + \sum_{j=1}^p \nu_j h_j(\boldsymbol{x})$.
- **Lagrange dual function**: $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) := \inf_{\boldsymbol{x}\in X} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$.
- **Weak Lagrange duality** ($\boldsymbol{\lambda} \geq 0$, $\boldsymbol{x}$ is feasible): $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f_0(\boldsymbol{x})$.
- **Lagrange dual problem** (convex program, even if primal is not):

$$\begin{aligned} &\text{maximize} && g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ &\text{subject to} && \boldsymbol{\lambda} \geq \boldsymbol{0}. \end{aligned}$$

- If a convex program has a feasible solution $\tilde{\boldsymbol{x}}$ that is a Slater point ($f_i(\tilde{\boldsymbol{x}}) < 0, \forall i \in [m]$), then $\max_{\boldsymbol{\lambda}\geq 0, \boldsymbol{\nu}} g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\boldsymbol{x}\in X} f_0(\boldsymbol{x})$.
- **Zero duality gap**: Feasible solutions $\tilde{\boldsymbol{x}}$ and $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ have zero duality gap if $f_0(\tilde{\boldsymbol{x}}) = g(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ ($\Rightarrow \tilde{\boldsymbol{x}}$ is a minimizer of primal).
- **KKT necessary**: Zero duality gap $\Rightarrow \tilde{\lambda} f_i(\tilde{\boldsymbol{x}}) = 0, \forall i \in [m]$ (complementary slackness) and $\nabla_{\boldsymbol{x}} L(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) = \boldsymbol{0}$ (vanishing Lagrangian gradient).
- **KKT sufficient**: Convex program, complementary slackness, and vanishing Lagrangian gradient $\Rightarrow$ Zero duality gap.

Complementary slackness ($f_0(\tilde{\boldsymbol{x}}) = L(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$) $\Rightarrow L$ is convex in $\boldsymbol{x}$ and gradient is zero, so $\tilde{\boldsymbol{x}}$ is a global minimizer. $\square$

- Program maybe not solvable, but if Slater point, then a solution exists $\Rightarrow$ Only need to show that the KKT conditions are satisfied.

Gradient descent

- **Update rule**: $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \gamma\nabla f(\boldsymbol{x}_t)$.
- **VA**: $\sum_{t=0}^{T-1}(f(\boldsymbol{x}_t) - f^\star) \leq \frac{\gamma}{2}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{1}{2\gamma}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2$.

1st-order convexity on $(\boldsymbol{x}^\star, \boldsymbol{x}_t) \Rightarrow \nabla f(\boldsymbol{x}_t) = \frac{\boldsymbol{x}_t - \boldsymbol{x}_{t+1}}{\gamma} \Rightarrow$ Cosine theorem $\Rightarrow \boldsymbol{x}_t - \boldsymbol{x}_{t+1} = \gamma\nabla f(\boldsymbol{x}_t) \Rightarrow$ Telescoping sum. $\square$

- **Sufficient decrease** (L-smooth, $\gamma := \frac{1}{L}$): $f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_t)\|^2$.

Smoothness on $(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t) \Rightarrow \boldsymbol{x}_{t+1} - \boldsymbol{x}_t = -\frac{1}{L}\nabla f(\boldsymbol{x}_t)$. $\square$

- **Convergence results**: ($\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\| \leq R$)
  - (B-Lipschitz, convex, $\gamma := \frac{R}{B\sqrt{T}}$) $\frac{1}{T}\sum_{t=0}^{T-1}(f(\boldsymbol{x}_t) - f^\star) \leq \frac{RB}{\sqrt{T}}$.

    Apply bounds to VA and find γ by 1st-order optimality. $\square$
  - (L-smooth, convex, $\gamma := \frac{1}{L}$) $f(\boldsymbol{x}_T) - f^\star \leq \frac{L}{2T}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2$

    Sufficient decrease to bound gradients of VA with telescoping sum. $\square$
  - (L-smooth, μ-SC, $\gamma := \frac{1}{L}$) $f(\boldsymbol{x}_T) - f^\star \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2$

    Use μ-SC to strengthen VA bound for squared norm $\Rightarrow$ Upper bound "noise" with $f^\star \leq f(\boldsymbol{x}_{t+1})$ and SD $\Rightarrow$ Smoothness on $(\boldsymbol{x}^\star, \boldsymbol{x}_T)$. $\square$
- **Accelerated gradient descent**:

$$\begin{aligned} \boldsymbol{y}_{t+1} &= \boldsymbol{x}_t - \frac{1}{L}\nabla f(\boldsymbol{x}_t) \\ \boldsymbol{z}_{t+1} &= \boldsymbol{z}_t - \frac{t+1}{2L}\nabla f(\boldsymbol{x}_t) \\ \boldsymbol{x}_{t+1} &= \frac{t+1}{t+3}\boldsymbol{y}_{t+1} + \frac{2}{t+3}\boldsymbol{z}_{t+1}. \end{aligned}$$

Projected gradient descent

- **Update rule** ($X \subset \mathbb{R}^d$ is closed and convex):

$$\begin{aligned} \boldsymbol{y}_{t+1} &= \boldsymbol{x}_t - \gamma\nabla f(\boldsymbol{x}_t) \\ \boldsymbol{x}_{t+1} &= \Pi_X(\boldsymbol{y}_{t+1}) := \operatorname*{argmin}_{\boldsymbol{x}\in X}\|\boldsymbol{x} - \boldsymbol{y}_{t+1}\|^2. \end{aligned}$$

- **Projection onto ℓ_1-ball** can be done in $\mathcal{O}(d\log d)$.

1. ($\boldsymbol{x} \in X, \boldsymbol{y} \in \mathbb{R}^d$): $\langle \boldsymbol{x} - \Pi_X(\boldsymbol{y}), \boldsymbol{y} - \Pi_X(\boldsymbol{y})\rangle \leq 0$.

   Constrained 1st-order optimality $\Rightarrow$ Rearrange. $\square$
2. ($\boldsymbol{x} \in X, \boldsymbol{y} \in \mathbb{R}^d$): $\|\boldsymbol{x} - \Pi_X(\boldsymbol{y})\|^2 + \|\boldsymbol{y} - \Pi_X(\boldsymbol{y})\|^2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|^2$.

   Cosine theorem on (1). $\square$

- If $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t$, then $\boldsymbol{x}_t = \boldsymbol{x}^\star$.

Use (1) and $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t$ to show that 1st-order optimality holds. $\square$

- **Projected SD**: $f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{L}{2}\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\|^2$.

Smoothness on $(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t) \Rightarrow \nabla f(\boldsymbol{x}_t) = L(\boldsymbol{y}_{t+1} - \boldsymbol{x}_t) \Rightarrow$ Cosine theorem $\Rightarrow \boldsymbol{y}_{t+1} - \boldsymbol{x}_t = -\frac{1}{L}\nabla f(\boldsymbol{x}_t)$. $\square$

- (L-smooth, convex, $\gamma := \frac{1}{L}$): $f(\boldsymbol{x}_T) - f^\star \leq \frac{L}{2T}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2$.

VA with additional term ($\boldsymbol{y}_{t+1}$ instead of $\boldsymbol{x}_{t+1}$ and use (2)) and bound gradients with projected SD. Additional terms cancel. $\square$

Coordinate descent

- **Update rule**: $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \gamma_i\nabla_i f(\boldsymbol{x}_t)\boldsymbol{e}_i, \quad i \in [d]$.
- **Coordinate-wise SD**: $f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \frac{1}{2L_i}|\nabla_i f(\boldsymbol{x}_t)|^2$.

CW smoothness with $\lambda = \frac{-\nabla_i f(\boldsymbol{x}_t)}{L_i}$ such that $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \lambda\boldsymbol{e}_i$. $\square$

- **Convergence results** (μ-PL, $\mathcal{L}$-CS, $\bar{L} = \frac{1}{d}\sum_{i=1}^d L_i$, $\gamma_i := \frac{1}{L_i}$):
  - (L-smooth, μ-PL, $i \sim \mathrm{Unif}([d])$)
    $\mathbb{E}[f(\boldsymbol{x}_T) - f^\star] \leq \left(1 - \frac{\mu}{dL}\right)^T(f(\boldsymbol{x}_0) - f^\star)$.

    CW SD $\Rightarrow \mathbb{E}_i[\cdot \mid \boldsymbol{x}_t] \Rightarrow$ Use sample prob. $\Rightarrow$ PL $\Rightarrow \mathbb{E}_{\boldsymbol{x}_t}$ (LoTE). $\square$
  - (μ-PL, $i \sim \mathrm{Cat}(L_1/\sum_{j=1}^d L_j, \dots, L_d/\sum_{j=1}^d L_j)$)
    $\mathbb{E}[f(\boldsymbol{x}_T) - f^\star] \leq \left(1 - \frac{\mu}{d\bar{L}}\right)^T(f(\boldsymbol{x}_0) - f^\star)$.

    Same as above with different probabilities. $\bar{L} := \frac{1}{d}\sum_{i=1}^d L_i$. $\square$
  - (L-smooth, μ_1-SC w.r.t. $\ell_1 \Rightarrow \mu_1$-PL w.r.t. ℓ_∞, $i \in \operatorname{argmax}_{j\in[d]}|\nabla_j f(\boldsymbol{x}_t)|$)
    $f(\boldsymbol{x}_T) - f^\star \leq \left(1 - \frac{\mu}{dL}\right)^T(f(\boldsymbol{x}_0) - f^\star)$
    $f(\boldsymbol{x}_T) - f^\star \leq \left(1 - \frac{\mu_1}{L}\right)^T(f(\boldsymbol{x}_0) - f^\star)$.

    CW SD $\Rightarrow \ell_\infty$ because of update rule $\Rightarrow$ PL. $\square$

    $\frac{1}{\sqrt{d}}\|\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|_1 \leq \|\boldsymbol{x} - \boldsymbol{y}\|_2 \Rightarrow \frac{\mu}{d} \leq \mu_1 \leq \mu$.

Nonconvex functions

- (L-smooth, $\gamma := \frac{1}{L}$, $\exists\boldsymbol{x}^\star$): $\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\|^2 \leq \frac{2L}{T}(f(\boldsymbol{x}_0) - f^\star)$.

SD does not require convexity. Rewrite with telescoping sum. $\square$

$\Rightarrow \lim_{t\to\infty}\|\nabla f(\boldsymbol{x}_t)\| = 0$.

- **Trajectory analysis**: Optimize $f(\boldsymbol{x}) := \frac{1}{2}\left(\prod_{k=1}^d x_k - 1\right)^2$.
- $\frac{\partial f(\boldsymbol{x})}{\partial x_i} = \left(\prod_k x_k - 1\right)\prod_{k\neq i} x_k$ ($\nabla f(\boldsymbol{x}) = \boldsymbol{0}$ if 2 dims are 0 or all 1).
- $\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i^2} = \left(\prod_{k\neq i} x_k\right)^2$.
- $\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i\,\partial x_j} = 2\prod_{k\neq i} x_k \prod_{k\neq j} x_k - \prod_{k\neq i,j} x_k$, if $i \neq j$.
- c**-balanced**: Let $\boldsymbol{x} > \boldsymbol{0}$, $c \geq 1$. $\boldsymbol{x}$ is c-balanced if $x_i \leq c \cdot x_j, \forall i, j \in [d]$.
- If $\boldsymbol{x}_t$ is c-balanced, $\gamma > 0$, then $\boldsymbol{x}_{t+1}$ is c-balanced and $\boldsymbol{x}_{t+1} \geq \boldsymbol{x}_t$.

- If $\boldsymbol{x}$ is c-balanced, then for any $I \subseteq [d]$, we have
$$\prod_{k \notin I} x_k \leq c^{|I|} \left(\prod_{k=1}^{d} x_k \right)^{1-|I|/d} \leq c^{|I|}.$$
- Let $\boldsymbol{x}$ be c-balanced and $\prod_k x_k \leq 1$, then
$$\|\nabla^2 f(\boldsymbol{x})\|_2 \leq \|\nabla^2 f(\boldsymbol{x})\|_F \leq 3dc^2.$$
Thus, f is smooth along the whole trajectory of GD with $L = 3dc^2$.
- **Convergence** ($\gamma := \frac{1}{3dc^2}$, $\boldsymbol{x}_0 > \boldsymbol{0}$ and c-balanced, $\delta \leq \prod_k x_{0,k} < 1$) $f(\boldsymbol{x}_T) \leq \left(1 - \frac{\delta^2}{3c^4}\right)^T f(\boldsymbol{x}_0)$.
- δ decays polynomially in d, so we must start $\mathcal{O}(1/\sqrt{d})$ from $\boldsymbol{x}^\star = \boldsymbol{1}$.

Frank-Wolfe

- **Linear minimization oracle**: $\mathrm{LMO}_X(\boldsymbol{g}) := \mathrm{argmin}_{\boldsymbol{z} \in X} \langle \boldsymbol{g}, \boldsymbol{z} \rangle$.
If $\boldsymbol{g} = \boldsymbol{0}$, any $\boldsymbol{z}$ minimizes.
- **Update rule**: $\boldsymbol{x}_{t+1} = (1-\gamma_t)\boldsymbol{x}_t + \gamma_t \boldsymbol{s}_t, \quad \boldsymbol{s}_t = \mathrm{LMO}_X(\nabla f(\boldsymbol{x}_t))$.
- If $X = \mathrm{conv}(\mathcal{A})$, then $\mathrm{LMO}_X(\boldsymbol{g}) \in \mathcal{A}$: Easy optimization problem in $\mathcal{O}(|\mathcal{A}|)$.
- Advantages: (1) Iterates are always feasible if X is convex, (2) No projections, (3) Iterates $\boldsymbol{x}_T$ have simple sparse representations as convex combination of $\{\boldsymbol{x}_0, \boldsymbol{s}_0, \ldots, \boldsymbol{s}_{T-1}\}$: $\boldsymbol{x}_T = \left(\prod_{t=0}^{T-1} 1 - \gamma_t\right)\boldsymbol{x}_0 + \sum_{t=0}^{T-1} \gamma_t \left(\prod_{\tau=t+1}^{T-1} 1 - \gamma_\tau\right)\boldsymbol{s}_t$.
- **ℓ_1-ball LMO**: $\mathrm{LMO}(\boldsymbol{g}) = -\mathrm{sgn}(g_i)\boldsymbol{e}_i, i \in \mathrm{argmax}_{j \in [d]} |g_j|$.
- **Spectahedron LMO**: $\mathrm{LMO}_X(\mathrm{G}) = \mathrm{argmin}_{\substack{\mathrm{tr}(Z)=1 \\ Z \text{ is PSD}}} G \odot Z = \boldsymbol{v}_1 \boldsymbol{v}_1^\top$, where $\boldsymbol{v}_1$ is the eigenvector associated with the smallest eigenvalue of G.
- **Duality gap**: $g(\boldsymbol{x}) := \langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{s} \rangle, \boldsymbol{s} = \mathrm{LMO}_X(\nabla f(\boldsymbol{x}))$.
- **Upper bound of optimality gap** (convex): $g(\boldsymbol{x}) \geq f(\boldsymbol{x}) - f^\star$.

> $g(\boldsymbol{x}) = \langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{s} \rangle \geq \langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^\star \rangle \geq f(\boldsymbol{x}) - f^\star$. $\square$

- **Descent lemma**: $f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \gamma_t g(\boldsymbol{x}_t) + \gamma_t^2 \frac{L}{2} \|\boldsymbol{s}_t - \boldsymbol{x}_t\|^2$.
- **Convergence** (L-smooth, convex, X is compact, $\gamma_t = \frac{2}{t+2}$): $f(\boldsymbol{x}_T) - f^\star \leq \frac{4C}{T+1}, \quad C = \frac{L}{2}\mathrm{diam}(X)^2$.

> Lemma$-f^\star$ $\Rightarrow$ Use $g(\boldsymbol{x}) \geq f(\boldsymbol{x}) - f^\star$ $\Rightarrow$ Rearrange and induction. $\square$

- **Affine equivalence**: (f, X) and (f', X') are affinely equivalent if $f'(\boldsymbol{x}) = f(A\boldsymbol{x} + \boldsymbol{b})$ and $X' = \{A^{-1}(\boldsymbol{x} - \boldsymbol{b}) \mid \boldsymbol{x} \in X\}$. Then,
$$\nabla f'(\boldsymbol{x}') = A^\top \nabla f(\boldsymbol{x}), \quad \boldsymbol{x}' = A^{-1}(\boldsymbol{x} - \boldsymbol{b})$$
$$\mathrm{LMO}_{X'}(\nabla f'(\boldsymbol{x}')) = A^{-1}(\boldsymbol{s} - \boldsymbol{b}), \quad \boldsymbol{s} = \mathrm{LMO}_X(\nabla f(\boldsymbol{x})).$$
- **Curvature constant**:
$$C_{(f,X)} := \sup_{\substack{\boldsymbol{x}, \boldsymbol{s} \in X, \gamma \in (0,1] \\ \boldsymbol{y} = (1-\gamma)\boldsymbol{x} + \gamma \boldsymbol{s}}} \frac{1}{\gamma^2}(f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle).$$
- **Affine invariant convergence** (same ass.): $f(\boldsymbol{x}_T) - f^\star \leq \frac{4C_{(f,X)}}{T+1}$.

> Descent lemma w.r.t. $C_{(f,X)}$ by setting $\boldsymbol{x} = \boldsymbol{x}_t, \boldsymbol{s} = \mathrm{LMO}_X(\nabla f(\boldsymbol{x}_t)), \gamma = \gamma_t, \boldsymbol{y} = \boldsymbol{x}_{t+1}$ in the supremum. Proof follows in the same way. $\square$

- **Convergence of** $g(\boldsymbol{x}_t)$: $\min_{1 \leq t \leq T} g(\boldsymbol{x}_t) \leq \frac{27/2 \cdot C_{(f,X)}}{T+1}$.

Newton's method

- **Update rule**: $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \nabla^2 f(\boldsymbol{x}_t)^{-1} \nabla f(\boldsymbol{x}_t)$.
- **Interp**: (1) Adaptive gradient descent, (2) Min. 2nd-order Taylor approx. at $\boldsymbol{x}_t$:
$$\boldsymbol{x}_{t+1} \in \underset{\boldsymbol{x} \in \mathbb{R}^d}{\mathrm{argmin}}\, f(\boldsymbol{x}_t) + \nabla f(\boldsymbol{x}_t)^\top (\boldsymbol{x} - \boldsymbol{x}_t) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_t)^\top \nabla^2 f(\boldsymbol{x}_t)(\boldsymbol{x} - \boldsymbol{x}_t).$$
- **Convergence** ($\|\nabla^2 f(\boldsymbol{x})^{-1}\| \leq \frac{1}{\mu}$, $\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \leq B\|\boldsymbol{x} - \boldsymbol{y}\|$): $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\| \leq \frac{B}{2\mu}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2$.

> $\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star \leq \boldsymbol{x}_t - \boldsymbol{x}^\star + H(\boldsymbol{x}_t)^{-1}(\nabla f(\boldsymbol{x}^\star) - \nabla f(\boldsymbol{x}_t)) \Rightarrow h(t) := \nabla f(\boldsymbol{x} + t(\boldsymbol{x}^\star - \boldsymbol{x}))$ with fundamental theorem of calculus $\Rightarrow$ Take norm of both sides and simplify using $\|A\boldsymbol{x}\| = \|A\|_2 \|\boldsymbol{x}\|$ and assumptions. $\square$

- Ensure bounded inverse Hessians by requiring strong convexity over X.
- If $\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\| \leq \frac{\mu}{B}$, then $\|\boldsymbol{x}_T - \boldsymbol{x}^\star\| \leq \frac{\mu}{B}\left(\frac{1}{2}\right)^{2^T - 1}$.

Quasi-Newton methods

- Time complexity of Hessian is $\mathcal{O}(d^3) \Rightarrow$ Approximate by H_t.
- **Secant condition**: $\nabla f(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}_{t-1}) = H_t(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})$.
- **Idea**: We wanted Hessian to fluctuate little in regions of fast convergence $\Rightarrow$ Update $H_t^{-1} = H_{t-1}^{-1} + E_t$ while minimizing $\|AEA^\top\|_F^2$ for some invertible A.
- $H := H_{t-1}^{-1}$, $H' := H_t^{-1}$, $E := E_t$, $\boldsymbol{\sigma} := \boldsymbol{x}_t - \boldsymbol{x}_{t-1}$, $\boldsymbol{y} := \nabla f(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}_{t-1})$, $\boldsymbol{r} := \boldsymbol{\sigma} - H\boldsymbol{y}$. Convex program:
$$\begin{array}{lll} \text{minimize} & \frac{1}{2}\|AEA^\top\|_F^2 & \\ \text{subject to} & E\boldsymbol{y} = \boldsymbol{r} & \text{(secant condition)} \\ & E^\top - E = 0. & \text{(symmetry)} \end{array}$$

- **Greenstadt method** ($\mathcal{O}(d^2)$): Solving (with Lagrange multipliers) yields
$$E^\star = \frac{1}{\boldsymbol{y}^\top M \boldsymbol{y}}\Big(\boldsymbol{\sigma}\boldsymbol{y}^\top M + M\boldsymbol{y}\boldsymbol{\sigma}^\top - H\boldsymbol{y}\boldsymbol{y}^\top M - M\boldsymbol{y}\boldsymbol{y}^\top H$$
$$- \frac{1}{\boldsymbol{y}^\top M \boldsymbol{y}}\Big(\boldsymbol{y}^\top \boldsymbol{\sigma} - \boldsymbol{y}^\top H \boldsymbol{y}\Big) M\boldsymbol{y}\boldsymbol{y}^\top M\Big)$$
for some matrix parameter M (induced by A).
- **BFGS**: Set $M = H'$: $E^\star = \frac{1}{\boldsymbol{y}^\top \boldsymbol{\sigma}}\Big(-H\boldsymbol{y}\boldsymbol{\sigma}^\top - \boldsymbol{\sigma}\boldsymbol{y}^\top H + \Big(1 + \frac{\boldsymbol{y}^\top H \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{\sigma}}\Big)\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\Big)$.
Equivalent update: $H' = \Big(I - \frac{\boldsymbol{\sigma}\boldsymbol{y}^\top}{\boldsymbol{y}^\top \boldsymbol{\sigma}}\Big) H \Big(I - \frac{\boldsymbol{y}\boldsymbol{\sigma}^\top}{\boldsymbol{y}^\top \boldsymbol{\sigma}}\Big) + \frac{\boldsymbol{\sigma}\boldsymbol{\sigma}^\top}{\boldsymbol{y}^\top \boldsymbol{\sigma}}$.
- **L-BFGS** ($\mathcal{O}(md)$): Recursive BFGS and only go down m steps.

Subgradient method

- Until now, we have only considered smooth (and hence differentiable) functions $\Rightarrow$ Generalize notion of gradient.
- **Update rule**: $\boldsymbol{x}_{t+1} = \Pi_X(\boldsymbol{x}_t - \gamma_t \boldsymbol{g}_t), \quad \boldsymbol{g}_t \in \partial f(\boldsymbol{x}_t)$.
- **Lemma** (convex): $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 \leq \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 - 2\gamma_t(f(\boldsymbol{x}_t) - f^\star) + \gamma_t^2\|\boldsymbol{g}_t\|^2$.

> Norm of update rule$-\boldsymbol{x}^\star$ $\Rightarrow$ Π_X is non-expansive $\Rightarrow$ Cosine theorem $\Rightarrow$ Subgradient definition on $(\boldsymbol{x}^\star, \boldsymbol{x}_t)$ (exists because of convexity). $\square$

- (convex): $\min_{1 \leq t \leq T} f(\boldsymbol{x}_t) - f^\star \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + \sum_{t=1}^T \gamma_t^2 \|\boldsymbol{g}_t\|^2}{2\sum_{t=1}^T \gamma_t}$.

> Rearrange "descent" lemma $\Rightarrow$ Sum and divide by $\sum_{t=1}^T \gamma_t$. $\square$

- (μ-SC, B-Lipschitz, $\gamma_t := \frac{2}{\mu(t+1)}$): $\min_{1 \leq t \leq T} f(\boldsymbol{x}_t) - f^\star \leq \frac{2B^2}{\mu(T+1)}$.

> Adapt "descent" lemma with μ-SC $\Rightarrow$ Def. of γ_t and $\|\boldsymbol{g}_t\| \leq B$. $\square$

Mirror descent

- Exploit non-Euclidean geometry of convex set X.
- **Bregman divergence**: Let $\omega : \Omega \to \mathbb{R}$ be continuously differentiable on Ω and 1-SC w.r.t. some norm $\|\cdot\|$. Then,
$$V_\omega(\boldsymbol{x}, \boldsymbol{y}) := \omega(\boldsymbol{x}) - \omega(\boldsymbol{y}) - \langle \nabla \omega(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle.$$
- **Properties**: $V_\omega(\boldsymbol{x}, \boldsymbol{y}) \geq 0$; $V_\omega(\boldsymbol{x}, \boldsymbol{y})$ is convex in $\boldsymbol{x}$; $V_\omega(\boldsymbol{x}, \boldsymbol{y}) = 0$ iff $\boldsymbol{x} = \boldsymbol{y}$; $V_\omega(\boldsymbol{x}, \boldsymbol{y}) \geq \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$; and $\nabla_{\boldsymbol{x}} V_\omega(\boldsymbol{x}, \boldsymbol{y}) = \nabla \omega(\boldsymbol{x}) - \nabla \omega(\boldsymbol{y})$.
- **3-point id.**: $V_\omega(\boldsymbol{x}, \boldsymbol{z}) = V_\omega(\boldsymbol{x}, \boldsymbol{y}) + V_\omega(\boldsymbol{y}, \boldsymbol{z}) - \langle \nabla \omega(\boldsymbol{z}) - \nabla \omega(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle$.
- **Update rule**: $\boldsymbol{x}_{t+1} \in \mathrm{argmin}_{\boldsymbol{x} \in X} V_\omega(\boldsymbol{x}, \boldsymbol{x}_t) + \langle \gamma_t \boldsymbol{g}_t, \boldsymbol{x} \rangle, \boldsymbol{g}_t \in \partial f(\boldsymbol{x}_t)$. This is a generalization of subgradient descent.
- **Lemma**: $\gamma_t(f(\boldsymbol{x}_t) - f^\star) \leq V_\omega(\boldsymbol{x}^\star, \boldsymbol{x}_t) - V_\omega(\boldsymbol{x}^\star, \boldsymbol{x}_{t+1}) + \frac{\gamma_t^2}{2}\|\boldsymbol{g}_t\|_\star^2$.

> Rearrange update rule constrained optimality condition $\Rightarrow$ 3PI $\Rightarrow$ $-V_\omega(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t) \leq -\frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2 \Rightarrow$ [Subgradient on $(\boldsymbol{x}^\star, \boldsymbol{x}_t)$] $\cdot \gamma_t$ ($\pm \boldsymbol{x}_{t+1}$ in inner product) and bound with prev. $\Rightarrow$ Young's inequality: $\langle \gamma_t \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle \leq \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2 + \frac{1}{2}\|\gamma_t \boldsymbol{g}_t\|_\star^2$. $\square$

- (Convex): $\min_{1 \leq t \leq T} f(\boldsymbol{x}_t) - f^\star \leq \frac{V_\omega(\boldsymbol{x}^\star, \boldsymbol{x}_0) + \frac{1}{2}\sum_{t=1}^T \gamma_t^2 \|\boldsymbol{g}_t\|_\star^2}{\sum_{t=1}^T \gamma_t}$.

> Easily follows from above lemma by summing, dividing by summed γ_t, and telescoping sum. $\square$

Smoothing

- **Nesterov smoothing**: $f_\mu(\boldsymbol{x}) := \max_{\boldsymbol{y} \in \mathrm{dom}(f^\star)} \langle \boldsymbol{x}, \boldsymbol{y} \rangle - f^\star(\boldsymbol{y}) - \mu \cdot d(\boldsymbol{y})$, where d is 1-SC and non-negative (proximity function).
- f_μ is $1/\mu$-smooth and approximates f by $f(\boldsymbol{x}) - \mu D^2 \leq f_\mu(\boldsymbol{x}) \leq f(\boldsymbol{x})$, $D^2 := \max_{\boldsymbol{y} \in \mathrm{dom}(f^\star)} d(\boldsymbol{y})$.
- Applying GD to f_μ converges faster than subgradient descent.
- **Moreau-Yosida smoothing**: $f_\mu(\boldsymbol{x}) := \min_{\boldsymbol{y} \in \mathrm{dom}(f^\star)} f(\boldsymbol{y}) - \frac{1}{2\mu}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$.
- f_μ is $1/\mu$-smooth and minimizes exactly: $\mathrm{argmin}_{\boldsymbol{x} \in X} f(\boldsymbol{x}) = \mathrm{argmin}_{\boldsymbol{x} \in X} f_\mu(\boldsymbol{x})$.
- $\nabla f_\mu(\boldsymbol{x}) = \frac{1}{\mu}(\boldsymbol{x} - \mathrm{prox}_{\mu f}(\boldsymbol{x}))$ (found by Danshkin's theorem).

Proximal algorithms

- **Proximal operator**: $\mathrm{prox}_{\mu f}(\boldsymbol{x}) := \mathrm{argmin}_{\boldsymbol{y} \in \mathrm{dom}(f)} f(\boldsymbol{y}) + \frac{1}{2\mu}\|\boldsymbol{x} - \boldsymbol{y}\|^2$.
- **Minimizer**: $\boldsymbol{x}^\star = \mathrm{prox}_{\mu f}(\boldsymbol{x}^\star), \quad \forall \mu$.
- **Non-expansiveness**: $\|\mathrm{prox}_{\mu f}(\boldsymbol{x}) - \mathrm{prox}_{\mu f}(\boldsymbol{y})\| \leq \|\boldsymbol{x} - \boldsymbol{y}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y}$.
- **Proximal point algorithm**: Apply gradient descent to Moreau-Yosida f_μ: $\boldsymbol{x}_{t+1} = \mathrm{prox}_{\lambda_t f}(\boldsymbol{x}_t)$.
- (Convex): $f(\boldsymbol{x}_{T+1}) - f^\star \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{2\sum_{t=1}^T \lambda_t}$

> Subgradient optimality: $-\frac{\boldsymbol{x}_{t+1} - \boldsymbol{x}_t}{\lambda_t} \in \partial f(\boldsymbol{x}_{t+1}) \Rightarrow$ Subgradient exists because of convexity $\Rightarrow$ Subgradient definition $\Rightarrow$ Cosine theorem $\Rightarrow$ Sum over timesteps and use that it is a descent method. $\square$

- **Proximal gradient method**: Consider $F(\boldsymbol{x}) := f(\boldsymbol{x}) + g(\boldsymbol{x})$ with differentiable f (both are convex): $\boldsymbol{x}_{t+1} = \mathrm{prox}_{\gamma_t g}(\boldsymbol{x}_t - \gamma_t \nabla f(\boldsymbol{x}_t))$.

- (f is L-smooth, $\gamma_t := \frac{1}{L}$): $F(\boldsymbol{x}_{T+1}) - F^\star \le \frac{L\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{2T}$.

Subgradient optimality: $\frac{1}{\gamma_t}(\boldsymbol{x}_t - \boldsymbol{x}_{t+1} - \gamma_t \nabla f(\boldsymbol{x}_t)) \in \partial g(\boldsymbol{x}_{t+1}) \Rightarrow$ Subgradient exists because of convexity $\Rightarrow$ Subgradient definition $\Rightarrow$ Cosine theorem $\Rightarrow$ $-\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x}\rangle = -\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{x}_t\rangle - \langle \nabla f(\boldsymbol{x}_{t+1}), \boldsymbol{x}_t - \boldsymbol{x}\rangle \Rightarrow$ Smoothness, convexity, and definition of γ_t. □

Stochastic optimization

- **Optimization problem**: $\min_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{\xi}}[f(\boldsymbol{x}, \boldsymbol{\xi})]$.
- **Unbiased gradient**: $\mathbb{E}_{\boldsymbol{\xi}}[\nabla f(\boldsymbol{x}, \boldsymbol{\xi}) \mid \boldsymbol{x}] = \nabla F(\boldsymbol{x})$ (typical assumption).
- **Update rule**: $\boldsymbol{\xi}_t \sim P$, $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \gamma_t \nabla f(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$.
- **Bounded variance**: $\mathbb{E}\big[\|\nabla f(\boldsymbol{x}_t, \boldsymbol{\xi}_t) - \nabla F(\boldsymbol{x})\|^2\big] \le \sigma^2$.
- (L-smooth, bounded variance, random output, $\gamma := \min\Big\{\frac{1}{L}, \frac{\gamma_0}{\sigma\sqrt{T}}\Big\}$): $\mathbb{E}\big[\|\nabla F(\hat{\boldsymbol{x}}_T)\|^2\big] \le \frac{\sigma}{\sqrt{T}}\Big(\frac{2(F(\boldsymbol{x}_1)-F^\star)}{\gamma_0} + L\gamma_0\Big) + \frac{2L(F(\boldsymbol{x}_1)-F^\star)}{T}$, where $\hat{\boldsymbol{x}}_T \sim \mathrm{Unif}(\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_T\})$.

Smoothness of F on $(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t)$ in $\mathbb{E} \Rightarrow$ Update rule: $\boldsymbol{x}_{t+1} - \boldsymbol{x}_t = -\gamma_t \nabla f(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$ $\Rightarrow \mathbb{E}[X^2] + \mathbb{E}[X]^2 + \mathrm{Var}[X]$: $\mathbb{E}\big[\|\nabla f(\boldsymbol{x}_t, \boldsymbol{\xi}_t)\|^2\big] = \|\nabla F(\boldsymbol{x}_t)\|^2 + \mathbb{E}\big[\|\nabla f(\boldsymbol{x}_t, \boldsymbol{\xi}_t) - \nabla F(\boldsymbol{x}_t)\|^2\big] \le \|\nabla F(\boldsymbol{x}_t)\|^2 + \sigma^2 \Rightarrow \gamma_t \le \frac{1}{L} \Rightarrow$ Rearrange $\Rightarrow$ Use definition of $\hat{\boldsymbol{x}}_T \Rightarrow$ Telescoping sum $\Rightarrow$ Definition of $\gamma_t \Rightarrow \max\{a, b\} \le a + b$ if $a, b \ge 0$. □

- (L-smooth, $\mathbb{E}\big[\|\nabla f(\boldsymbol{x}, \boldsymbol{\xi})\|^2\big] \le B^2$): $\mathbb{E}[F(\hat{\boldsymbol{x}}_T) - F^\star] \le \frac{R^2 + B^2\sum_{t=1}^T \gamma_t^2}{2\sum_{t=1}^T \gamma_t}$, where $\hat{\boldsymbol{x}_t} := \frac{\sum_{t=1}^T \gamma_t \boldsymbol{x}_t}{\sum_{t=1}^T \gamma_t}$ and $\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\| \le R$.

Squared norm of update rule$-\boldsymbol{x}^\star \Rightarrow$ Cosine theorem $\Rightarrow$ Law of total exp. to bound inner product $\Rightarrow$ Convexity of $F \Rightarrow$ Telescoping sum $\Rightarrow$ Jensen's ineq. □

- (μ-SC, $\mathbb{E}\big[\|\nabla f(\boldsymbol{x}, \boldsymbol{\xi})\|^2\big] \le B^2$, $\gamma_t := \frac{\gamma}{t}$, $\gamma > \frac{1}{2\mu}$) $\mathbb{E}\big[\|\boldsymbol{x}_T - \boldsymbol{x}^\star\|^2\big] \le \frac{\max\{\frac{\gamma^2 B^2}{2\mu\gamma - 1}, \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2\}}{T}$.

Squared norm of update rule$-\boldsymbol{x}^\star \Rightarrow$ Cosine theorem $\Rightarrow$ μ-SC to get $\mathbb{E}[\langle \nabla f(\boldsymbol{x}_t, \boldsymbol{\xi}_t), \boldsymbol{x}_t - \boldsymbol{x}^\star\rangle] \ge \mu \cdot \mathbb{E}\big[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2\big] \Rightarrow$ Recursion. □

- **Adaptive method**: $\boldsymbol{g}_t = \nabla f(\boldsymbol{x}_t, \boldsymbol{\xi}_t)$, $\boldsymbol{m}_t = \phi_t(\boldsymbol{g}_1, \dots, \boldsymbol{g}_t)$, $V_t = \psi_t(\boldsymbol{g}_1, \dots, \boldsymbol{g}_t)$, $\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t - \alpha_t V_t^{-1/2}\boldsymbol{m}_t$, $\boldsymbol{x}_{t+1} = \mathrm{argmin}_{\boldsymbol{x}\in X}\Big\{(\boldsymbol{x} - \hat{\boldsymbol{x}}_t)^\top V_t^{-1/2}(\boldsymbol{x} - \hat{\boldsymbol{x}}_t)\Big\}$.
  - **SGD**: $\boldsymbol{m}_t = \boldsymbol{g}_t$, $V_t = I$.
  - **AdaGrad**: $\boldsymbol{m}_t = \boldsymbol{g}_t$, $V_t = \frac{\mathrm{diag}(\sum_{\tau=1}^t \boldsymbol{g}_\tau^2)}{t}$.
  - **Adam**: $\boldsymbol{m}_t = (1-\alpha)\sum_{\tau=1}^t \alpha^{t-\tau}\boldsymbol{g}_\tau$, $V_t = (1-\beta)\mathrm{diag}\big(\sum_{\tau=1}^t \beta^{t-\tau}\boldsymbol{g}_\tau^2\big)$. Recursively: $\boldsymbol{m}_t = \alpha\boldsymbol{m}_{t-1} + (1-\alpha)\boldsymbol{g}_t$, $V_t = \beta V_{t-1} + (1-\beta)\mathrm{diag}(\boldsymbol{g}_t^2)$.

Variance reduction

- SGD requires more iterations due to high variance $\Rightarrow$ Reduce variance.
- **Finite-sum optimization**: $\min_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^n f_i(\boldsymbol{x})$.
- If we want to estimate $\theta = \mathbb{E}[X]$, we can also estimate $\boldsymbol{\theta}$ as $\mathbb{E}[X - Y]$ if and only if $\mathbb{E}[Y] = 0$. Furthermore, $\mathrm{Var}[X - Y] \le \mathrm{Var}[X]$ if Y is highly positively correlated with X. Specifically, if $\mathrm{Cov}(X, Y) > \frac{1}{2}\mathrm{Var}[Y]$, the variance will be reduced.
- Let $\alpha \in [0, 1]$, we estimate θ by $\hat{\theta}_\alpha = \alpha(X - Y) + \mathbb{E}[Y]$. We then have
$$\mathbb{E}[\hat{\theta}_\alpha] = \alpha\mathbb{E}[X] + (1-\alpha)\mathbb{E}[Y]$$
$$\mathrm{Var}[\hat{\theta}_\alpha] = \alpha^2(\mathrm{Var}[X] + \mathrm{Var}[Y] - 2\cdot\mathrm{Cov}(X, Y)).$$
Implication: Trade-off between bias and variance, where $\alpha = 1$ makes the estimator unbiased, but the variance decreases when α decreases.
- SGD estimates $\nabla F(\boldsymbol{x}_t)$ by $\nabla f_{i_t}(\boldsymbol{x}_t)$, but VR estimates the full gradient by
$$\boldsymbol{g}_t := \alpha(\nabla f_{i_t}(\boldsymbol{x}_t) - Y) + \mathbb{E}[Y],$$
such that $\boldsymbol{g}_t$ satisfies the **VR property**: $\lim_{t\to\infty} \mathbb{E}\big[\|\boldsymbol{g}_t - \nabla F(\boldsymbol{x}_t)\|^2\big] = 0$.
- **Key idea**: If $\boldsymbol{x}_t$ is not too far away from previous iterates $\boldsymbol{x}_{1:t-1}$, we can leverage previous gradient information to construct positively correlated control variates Y.
  - **Stochastic Average Gradient (SAG)**: Keep track of the latest gradients $\boldsymbol{v}_i^t$ for all points $i \in [n]$: $\mathcal{O}(nd)$ storage requirement. Estimate full gradient by average of these: $\boldsymbol{g}_t = \frac{1}{n}\sum_{i=1}^n \boldsymbol{v}_i^t$. Each iteration we update $\boldsymbol{v}_i^t$ by
$$\boldsymbol{v}_i^t = \begin{cases} \nabla f_{i_t}(\boldsymbol{x}_t) & i = i_t \\ \boldsymbol{v}_i^{t-1} & i \ne i_t. \end{cases}$$
Thus, we have $\alpha = \frac{1}{n}$, $Y = \boldsymbol{v}_{i_t}^{t-1}$, and $\mathbb{E}[Y] = \boldsymbol{g}_{t-1}$,
$$\boldsymbol{g}_t = \frac{1}{n}\Big(\nabla f_{i_t}(\boldsymbol{x}_t) - \boldsymbol{v}_{i_t}^{t-1}\Big) + \boldsymbol{g}_{t-1}.$$
Problem: (1) $\mathcal{O}(nd)$ storage, (2) biased $\alpha \ne 1$. Advantage: $\mathcal{O}((n + \kappa_{\max}\log\frac{1}{\epsilon}))$ iteration complexity, where $\kappa_{\max} = \max_{i\in[n]} \frac{L_i}{\mu}$.
  - **SAGA**: Unbiased version of SAG, because it sets $\alpha = 1$: $\boldsymbol{g}_t = \nabla f_{i_t}(\boldsymbol{x}_t) - \boldsymbol{v}_{i_t}^{t-1} + \boldsymbol{g}_{t-1}$. But, it still enjoys the same benefits.
  - **Stochastic variance reduced gradient (SVRG)**: Build covariates based on a fixed reference point $\tilde{\boldsymbol{x}}$ that is periodically updated every m-th iteration:
$$\boldsymbol{g}_t = \nabla f_{i_t}(\boldsymbol{x}_t) - \nabla f_{i_t}(\tilde{\boldsymbol{x}}) + \nabla F(\tilde{\boldsymbol{x}}).$$
Problems: (1) $\mathcal{O}(n + 2m)$ gradient evaluations per epoch, (2) More hyperparameters. Advantages: (1) Unbiased, (2) $\mathcal{O}(d)$ memory cost, (3) Same iteration complexity as SAG(A).

Min-max optimization

- **Optimization problem**: $\min_{\boldsymbol{x}\in X}\max_{\boldsymbol{y}\in Y} \phi(\boldsymbol{x}, \boldsymbol{y})$.
- **Saddle point**: $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ is a saddle point if
$$\phi(\boldsymbol{x}^\star, \boldsymbol{y}) \le \phi(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \le \phi(\boldsymbol{x}, \boldsymbol{y}^\star), \quad \forall \boldsymbol{x} \in X, \boldsymbol{y} \in Y.$$
Interpretation: No player has the incentive to make a unilateral change, because it can only get worse. Game theory: Nash equilibrium.
- **Global minimax point**: $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ is a global minimax point if
$$\phi(\boldsymbol{x}^\star, \boldsymbol{y}) \le \phi(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \le \max_{\boldsymbol{y}'\in Y} \phi(\boldsymbol{x}, \boldsymbol{y}'), \quad \forall \boldsymbol{x} \in X, \boldsymbol{y} \in Y.$$
Interpretation: $\boldsymbol{x}^\star$ is the best response to the best response. Game theory: Stackelberg equilibrium.
- $\max_{\boldsymbol{y}\in Y}\min_{\boldsymbol{x}\in X} \phi(\boldsymbol{x}, \boldsymbol{y}) \le \min_{\boldsymbol{x}\in X}\max_{\boldsymbol{y}\in Y} \phi(\boldsymbol{x}, \boldsymbol{y})$.
- **Saddle point lemma**: $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ is a saddle point iff $\max_{\boldsymbol{y}\in Y}\min_{\boldsymbol{x}\in X} \phi(\boldsymbol{x}, \boldsymbol{y}) = \min_{\boldsymbol{x}\in X}\max_{\boldsymbol{y}\in Y} \phi(\boldsymbol{x}, \boldsymbol{y})$ and $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ are the arguments.
- **Minimax theorem**: If X and Y are closed convex sets, one of them is bounded, and ϕ is a continuous C-C function, then there exists a saddle point in $X \times Y$.
- **Duality gap**: $\hat{\epsilon}(\boldsymbol{x}, \boldsymbol{y}) := \max_{\boldsymbol{y}'\in Y} \phi(\boldsymbol{x}, \boldsymbol{y}') - \min_{\boldsymbol{x}'\in X} \phi(\boldsymbol{x}', \boldsymbol{y}) \ge 0$.
- **Saddle point by duality gap**: If $\hat{\epsilon}(\boldsymbol{x}, \boldsymbol{y}) = 0$, then $(\boldsymbol{x}, \boldsymbol{y})$ is a saddle point and if $\hat{\epsilon}(\boldsymbol{x}, \boldsymbol{y}) \le \epsilon$, then $(\boldsymbol{x}, \boldsymbol{y})$ is an ϵ-saddle point.
- **Gradient descent ascent (GDA)**: $\boldsymbol{x}_{t+1} = \Pi_X(\boldsymbol{x}_t - \gamma\nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}_t, \boldsymbol{y}_t))$, $\boldsymbol{y}_{t+1} = \Pi_Y(\boldsymbol{y}_t + \gamma\nabla_{\boldsymbol{y}}\phi(\boldsymbol{x}_t, \boldsymbol{y}_t))$. Does not guarantee convergence in C-C setting (consider $\phi(x, y) = xy$).
- (L-smooth, μ-SC-SC, $\gamma := \frac{\mu}{4L^2}$): $\|\boldsymbol{x}_T - \boldsymbol{x}^\star\|^2 + \|\boldsymbol{y}_T - \boldsymbol{y}^\star\|^2 \le \Big(1 - \frac{\mu^2}{4L^2}\Big)^T(\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + \|\boldsymbol{y}_1 - \boldsymbol{y}^\star\|^2)$.

Add μ-SC-SC definitions together $\Rightarrow$ Use L-smoothness for a bound $\Rightarrow$ Use update rule in $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 + \|\boldsymbol{y}_{t+1} - \boldsymbol{y}^\star\|^2 \Rightarrow$ Non-expansiveness of projection $\Rightarrow$ Rearrange $\Rightarrow$ Cosine theorem $\Rightarrow$ Bound inner products using SC-SC and smoothness. □

- **Extragradient method (EG)**:
$$\boldsymbol{x}_{t+1/2} = \Pi_X(\boldsymbol{x}_t - \gamma\nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}_t, \boldsymbol{y}_t))$$
$$\boldsymbol{y}_{t+1/2} = \Pi_Y(\boldsymbol{y}_t + \gamma\nabla_{\boldsymbol{y}}\phi(\boldsymbol{x}_t, \boldsymbol{y}_t))$$
$$\boldsymbol{x}_{t+1} = \Pi_X(\boldsymbol{x}_t - \gamma\nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}_{t+1/2}, \boldsymbol{y}_{t+1/2}))$$
$$\boldsymbol{y}_{t+1} = \Pi_Y(\boldsymbol{y}_t + \gamma\nabla_{\boldsymbol{y}}\phi(\boldsymbol{x}_{t+1/2}, \boldsymbol{y}_{t+1/2})).$$
- (L-smooth, C-C, $\gamma \le \frac{1}{2L}$): $\hat{\epsilon}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \le \frac{D_X^2 + D_Y^2}{2\gamma T}$, where $\bar{\boldsymbol{x}} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_{t+1/2}$, $\bar{\boldsymbol{y}} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{y}_{t+1/2}$, and $D_Z = \max_{\boldsymbol{z}, \boldsymbol{z}'\in Z}\|\boldsymbol{z} - \boldsymbol{z}'\|$.
- (L-smooth, μ-SC-SC, $\gamma := \frac{1}{8L}$): $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 + \|\boldsymbol{y}_{t+1} - \boldsymbol{y}^\star\|^2 \le \big(1 - \frac{\mu}{4L}\big)\big(\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + \|\boldsymbol{y}_t - \boldsymbol{y}^\star\|^2\big)$.
- **Optimistic gradient descent ascent (OGDA)**:
$$\boldsymbol{x}_{t+1/2} = \Pi_X(\boldsymbol{x}_t - \gamma\nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}_{t-1/2}, \boldsymbol{y}_{t-1/2}))$$
$$\boldsymbol{y}_{t+1/2} = \Pi_Y(\boldsymbol{y}_t + \gamma\nabla_{\boldsymbol{y}}\phi(\boldsymbol{x}_{t-1/2}, \boldsymbol{y}_{t-1/2}))$$
$$\boldsymbol{x}_{t+1} = \Pi_X(\boldsymbol{x}_t - \gamma\nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}_{t+1/2}, \boldsymbol{y}_{t+1/2}))$$
$$\boldsymbol{y}_{t+1} = \Pi_Y(\boldsymbol{y}_t + \gamma\nabla_{\boldsymbol{y}}\phi(\boldsymbol{x}_{t+1/2}, \boldsymbol{y}_{t+1/2})).$$
- In the case $X = Y = \mathbb{R}^d$, this can be seen as negative momentum:
$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - 2\gamma\nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}_t, \boldsymbol{y}_t) + \gamma\nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})$$
$$\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + 2\gamma\nabla_{\boldsymbol{y}}\phi(\boldsymbol{x}_t, \boldsymbol{y}_t) - \gamma\nabla_{\boldsymbol{y}}\phi(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}).$$
- **Proximal point algorithm**:
$$(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1}) \in \operatorname*{argmin}_{\boldsymbol{x}\in X}\operatorname*{argmax}_{\boldsymbol{y}\in Y} \phi(\boldsymbol{x}, \boldsymbol{y}) + \frac{1}{2\gamma}\|\boldsymbol{x} - \boldsymbol{x}_t\|^2 - \frac{1}{2\gamma}\|\boldsymbol{y} - \boldsymbol{y}_t\|^2.$$

Variational inequalities

- Generalizes all of the above to mapping $F : \mathcal{Z} \to \mathbb{R}^d$. Goal: Find $\boldsymbol{z}^\star \in \mathcal{Z}$, such that $\langle F(\boldsymbol{z}^\star), \boldsymbol{z} - \boldsymbol{z}^\star\rangle \ge 0, \forall \boldsymbol{z} \in \mathcal{Z}$.
- **Monotone operator**: $\langle F(\boldsymbol{x}) - F(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \ge 0$.
- **μ-strongly monotone**: $\langle F(\boldsymbol{x}) - F(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \ge \mu\|\boldsymbol{x} - \boldsymbol{y}\|^2$.
- **VI strong solution (Stampacchia)**: $\langle F(\boldsymbol{z}^\star), \boldsymbol{z} - \boldsymbol{z}^\star\rangle \ge 0, \forall \boldsymbol{z} \in \mathcal{Z}$.
- **VI weak solution (Minty)**: $\langle F(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{z}^\star\rangle \ge 0, \forall \boldsymbol{z} \in \mathcal{Z}$.
- If F is monotone, then strong $\Rightarrow$ weak. If F is continuous, then weak $\Rightarrow$ strong.
- Convex minimization can be cast as VI problem by defining $F = \nabla f$ for a convex function. Min-max problems can be cast as VI problem by defining $F = [\nabla_{\boldsymbol{x}}\phi, -\nabla_{\boldsymbol{y}}\phi]$ for a convex-concave ϕ.
- **Extragradient method**:
$$\boldsymbol{z}_{t+1/2} = \Pi_{\mathcal{Z}}(\boldsymbol{z}_t - \gamma_t F(\boldsymbol{z}_t))$$
$$\boldsymbol{z}_{t+1} = \Pi_{\mathcal{Z}}(\boldsymbol{z}_t - \gamma_t F(\boldsymbol{z}_{t+1/2})).$$
- (L-smooth, monotone, $\gamma := \frac{1}{\sqrt{2}L}$): $\max_{\boldsymbol{z}\in\mathcal{Z}}\langle F(\boldsymbol{z}), \bar{\boldsymbol{z}} - \boldsymbol{z}\rangle \le \frac{\sqrt{2}LD_{\mathcal{Z}}^2}{T}$, where $\bar{\boldsymbol{z}} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{z}_{t+1/2}$.

Optimality condition w.r.t. $\boldsymbol{z}_{t+1/2} \Rightarrow$ Rewrite using cosine theorem $\Rightarrow$ Optimality condition w.r.t. $\boldsymbol{z}_{t+1}$ (set $\boldsymbol{z} = \boldsymbol{z}_{t+1}$ in the other optimality condition) $\Rightarrow$ Use previous and Cauchy-Schwarz to bound $2\gamma\langle F(\boldsymbol{z}_{t+1/2}), \boldsymbol{z}_{t+1/2} - \boldsymbol{z}\rangle = 2\gamma\langle F(\boldsymbol{z}_{t+1/2}), \boldsymbol{z}_{t+1/2} - \boldsymbol{z}_{t+1}\rangle + 2\gamma\langle F(\boldsymbol{z}_{t+1/2}), \boldsymbol{z}_{t+1} - \boldsymbol{z}\rangle \Rightarrow$ Smoothness and $\gamma = \frac{1}{L} \Rightarrow$ Young's inequality: $\|\boldsymbol{x}\|\cdot\|\boldsymbol{y}\| \le \frac{1}{2}\|\boldsymbol{x}\|^2 + \frac{1}{2}\|\boldsymbol{y}\|^2 \Rightarrow$ Use monotonicity and sum over all timesteps. □