

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

Definitions

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

Lemmas

- $\frac{d}{dx} \frac{f(x)}{g(x)} = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$.
- Cosine theorem:** All equivalent formulations,
 - $\|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\langle \mathbf{x}, \mathbf{y} \rangle$.
 - $\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{2}(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2)$.
 - $\langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{z} \rangle = \frac{1}{2}(\|\mathbf{x} - \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{z}\|^2 - \|\mathbf{y} - \mathbf{z}\|^2)$.
- Cauchy-Schwarz:**
 - $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|$.
 - $(\sum_{i=1}^n a_i b_i)^2 \leq (\sum_{i=1}^n a_i^2)(\sum_{i=1}^n b_i^2)$.
 - Titu's lemma** ($b_i \geq 0$): $\frac{(\sum_{i=1}^n a_i)^2}{\sum_{i=1}^n b_i} \leq \sum_{i=1}^n \frac{a_i^2}{b_i}$ (proof: $a'_i = \frac{a_i}{\sqrt{b_i}}, b'_i = \sqrt{b_i}$).
- Hölder's inequality** (special case): $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|_1 \|\mathbf{y}\|_\infty$.
- Parallelogram law:** $2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2 = \|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2$.
- Jensen's inequality** (φ convex, $a_i \geq 0$): $\varphi\left(\frac{\sum_{i=1}^m a_i \mathbf{x}_i}{\sum_{i=1}^m a_i}\right) \leq \frac{\sum_{i=1}^m a_i \varphi(\mathbf{x}_i)}{\sum_{i=1}^m a_i}$.
- Fenchel's inequality:** $\langle \mathbf{x}, \mathbf{y} \rangle \leq f(\mathbf{x}) + f^*(\mathbf{y}) \Rightarrow \langle \mathbf{x}, \mathbf{y} \rangle \leq \frac{1}{2}(\|\mathbf{x}\|^2 + \|\mathbf{y}\|_*^2)$.
- Young's inequality** ($a, b \geq 0, \frac{1}{p} + \frac{1}{q} = 1$): $ab \leq \frac{a^p}{p} + \frac{b^q}{q} \Rightarrow \|\mathbf{x}\| \|\mathbf{y}\| \leq \frac{1}{2}(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2)$.
- $\frac{1}{\sqrt{d}}\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{d}\|\mathbf{x}\|_2$.
- $\|A\mathbf{x}\| \leq \|A\|_2 \|\mathbf{x}\|$.
- $\|A\|_2 \leq \|A\|_F$.

- 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 $[a, b]$):

$$h(b) - h(a) = \int_a^b h'(t) dt.$$

- $\left\| \int_0^1 \nabla h(t) dt \right\| \leq \int_0^1 \|\nabla h(t)\| dt$.

- $\int_0^1 c dt = c, \quad \int_0^1 t dt = \frac{1}{2}$.

- Subgradient calculus:**

- $h(\mathbf{x}) = \alpha f(\mathbf{x}) + \beta g(\mathbf{x}) \Rightarrow \partial h(\mathbf{x}) = \alpha \cdot \partial f(\mathbf{x}) + \beta \cdot \partial g(\mathbf{x})$.
 - $h(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b}) \Rightarrow \partial h(\mathbf{x}) = A^\top \partial f(A\mathbf{x} + \mathbf{b})$.
 - $h(\mathbf{x}) = \max f_i(\mathbf{x}) \Rightarrow \partial h(\mathbf{x}) = \text{conv}(\{\partial f_i(\mathbf{x}) \mid f_i(\mathbf{x}) = h(\mathbf{x})\})$.
- If f is differentiable at \mathbf{x} , then $\partial f(\mathbf{x}) \subseteq \{\nabla f(\mathbf{x})\}$.
 - If f is convex, then $\partial f(\mathbf{x}) \neq \emptyset$ for all \mathbf{x} in the relative interior.
 - If $\text{dom}(f)$ convex and $\partial f(\mathbf{x}) \neq \emptyset, \forall \mathbf{x} \in \text{dom}(f)$, 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, $\|\cdot\|_{p,*} = \|\cdot\|_q$.

Optimality lemmas (assume convexity)

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

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

Common tricks

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

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:
 - State what needs to be shown exactly and mark by $(*)$.
 - State the assumptions of the question and their implications (think about which implications are relevant to the proof).
 - Proof should follow easily: "Hence, $(*)$ 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

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

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 $1 - \delta$.

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

- 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 $\ell_n(H_n) \leq \inf_{H \in \mathcal{H}} \ell_n(H) + \epsilon \leq \ell(\tilde{H}) + 2\epsilon \leq \ell(\tilde{H}_n) + 3\epsilon$.

Non-linear programming

- Optimization problem:**

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

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

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

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

1st-order convexity on $(\mathbf{x}^*, \mathbf{x}_t) \Rightarrow \nabla f(\mathbf{x}_t) = \frac{\mathbf{x}_t - \mathbf{x}_{t+1}}{\gamma} \Rightarrow$ Cosine theorem $\Rightarrow \mathbf{x}_t - \mathbf{x}_{t+1} = \gamma \nabla f(\mathbf{x}_t) \Rightarrow$ Telescoping sum. \square
- Sufficient decrease** (L -smooth, $\gamma := \frac{1}{L}$): $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|^2$.

Smoothness on $(\mathbf{x}_{t+1}, \mathbf{x}_t) \Rightarrow \mathbf{x}_{t+1} - \mathbf{x}_t = -\frac{1}{L} \nabla f(\mathbf{x}_t)$. \square
- Convergence results:** ($\|\mathbf{x}_0 - \mathbf{x}^*\| \leq R$)
 - (B -Lipschitz, convex, $\gamma := \frac{R}{B\sqrt{T}}$) $\frac{1}{T} \sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f^*) \leq \frac{RB}{\sqrt{T}}$.

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

Sufficient decrease to bound gradients of VA with telescoping sum. \square

$$\circ (L\text{-smooth, } \mu\text{-SC, } \gamma := \frac{1}{L}) f(\mathbf{x}_T) - f^* \leq \frac{L}{2} \left(1 - \frac{\mu}{L}\right)^T \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

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

Accelerated gradient descent:

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

Projected gradient descent

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

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t)$$

$$\mathbf{x}_{t+1} = \Pi_X(\mathbf{y}_{t+1}) := \operatorname{argmin}_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{y}_{t+1}\|^2.$$
- Projection onto ℓ_1 -ball** can be done in $\mathcal{O}(d \log d)$.
- 1. $(\mathbf{x} \in X, \mathbf{y} \in \mathbb{R}^d): \langle \mathbf{x} - \Pi_X(\mathbf{y}), \mathbf{y} - \Pi_X(\mathbf{y}) \rangle \leq 0$.

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

Cosine theorem on (1). \square
- If $\mathbf{x}_{t+1} = \mathbf{x}_t$, then $\mathbf{x}_t = \mathbf{x}^*$.

Use (1) and $\mathbf{x}_{t+1} = \mathbf{x}_t$ to show that 1st-order optimality holds. \square
- Projected SD:** $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2} \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2$.

Smoothness on $(\mathbf{x}_{t+1}, \mathbf{x}_t) \Rightarrow \nabla f(\mathbf{x}_t) = L(\mathbf{y}_{t+1} - \mathbf{x}_t) \Rightarrow$ Cosine theorem $\Rightarrow \mathbf{y}_{t+1} - \mathbf{x}_t = -\frac{1}{L} \nabla f(\mathbf{x}_t)$. \square
- (L -smooth, convex, $\gamma := \frac{1}{L}$): $f(\mathbf{x}_T) - f^* \leq \frac{L}{2T} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$.

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

Coordinate descent

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

CW smoothness with $\lambda = \frac{-\nabla_i f(\mathbf{x}_t)}{L_i}$ such that $\mathbf{x}_{t+1} = \mathbf{x}_t + \lambda \mathbf{e}_i$. \square
- Convergence results** (μ -PL, L -CS, $\bar{L} = \frac{1}{d} \sum_{i=1}^d L_i, \gamma_i := \frac{1}{L_i}$):
 - (L -smooth, μ -PL, $i \sim \text{Unif}([d])$)

$$\mathbb{E}[f(\mathbf{x}_T) - f^*] \leq \left(1 - \frac{\mu}{dL}\right)^T (f(\mathbf{x}_0) - f^*).$$

CW SD $\Rightarrow \mathbb{E}_i[\cdot | \mathbf{x}_t] \Rightarrow$ Use sample prob. \Rightarrow PL $\Rightarrow \mathbb{E}_{\mathbf{x}_t}$ (LoTE). \square
 - (μ -PL, $i \sim \text{Cat}(L_1/\sum_{j=1}^d L_j, \dots, L_d/\sum_{j=1}^d L_j)$)

$$\mathbb{E}[f(\mathbf{x}_T) - f^*] \leq \left(1 - \frac{\mu}{dL}\right)^T (f(\mathbf{x}_0) - f^*).$$

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. $\ell_\infty, i \in \operatorname{argmax}_{j \in [d]} \|\nabla_j f(\mathbf{x}_t)\|$)

$$f(\mathbf{x}_T) - f^* \leq \left(1 - \frac{\mu_1}{dL}\right)^T (f(\mathbf{x}_0) - f^*)$$

$$f(\mathbf{x}_T) - f^* \leq \left(1 - \frac{\mu_1}{L}\right)^T (f(\mathbf{x}_0) - f^*).$$

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

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

Nonconvex functions

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

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

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

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

Frank-Wolfe

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

$$g(\mathbf{x}) = \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{s} \rangle \geq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^* \rangle \geq f(\mathbf{x}) - f^*. \quad \square$$

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

$$\text{Lemma } -f^* \Rightarrow \text{Use } g(\mathbf{x}) \geq f(\mathbf{x}) - f^* \Rightarrow \text{Rearrange and induction.} \quad \square$$

- Affine equivalence:** (f, X) and (f', X') are affinely equivalent if $f'(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b})$ and $X' = \{A^{-1}(\mathbf{x} - \mathbf{b}) \mid \mathbf{x} \in X\}$. Then, $\nabla f'(\mathbf{x}') = A^\top \nabla f(\mathbf{x}), \mathbf{x}' = A^{-1}(\mathbf{x} - \mathbf{b})$. $\text{LMO}_{X'}(\nabla f'(\mathbf{x}')) = A^{-1}(\mathbf{s} - \mathbf{b}), \mathbf{s} = \text{LMO}_X(\nabla f(\mathbf{x}))$.

- Curvature constant:** $C_{(f,X)} := \sup_{\substack{\mathbf{x}, \mathbf{s} \in X, \gamma \in (0,1] \\ \mathbf{y} = (1-\gamma)\mathbf{x} + \gamma\mathbf{s}}} \frac{1}{\gamma^2} (f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle)$.

- Affine invariant convergence** (same ass.): $f(\mathbf{x}_T) - f^* \leq \frac{4C_{(f,X)}}{T+1}$.

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

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

Newton's method

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

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

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

Quasi-Newton methods

- Time complexity of Hessian is $\mathcal{O}(d^3) \Rightarrow$ Approximate by H_t .
- Secant condition:** $\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}) = H_t(\mathbf{x}_t - \mathbf{x}_{t-1})$.
- 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 .
- $H := H_{t-1}^{-1}, H' := H_t^{-1}, E := E_t, \sigma := \mathbf{x}_t - \mathbf{x}_{t-1}, \mathbf{y} := \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}), \mathbf{r} := \sigma - H\mathbf{y}$. Convex program:

$$\begin{aligned} &\text{minimize} && \frac{1}{2} \|AEA^\top\|_F^2 \\ &\text{subject to} && E\mathbf{y} = \mathbf{r} \quad (\text{secant condition}) \\ &&& E^\top - E = 0. \quad (\text{symmetry}) \end{aligned}$$

Greenstadt method ($\mathcal{O}(d^2)$): Solving (with Lagrange multipliers) yields

$$E^* = \frac{1}{\mathbf{y}^\top M \mathbf{y}} \left(\sigma \mathbf{y}^\top M + M \mathbf{y} \sigma^\top - H \mathbf{y} \mathbf{y}^\top M - M \mathbf{y} \mathbf{y}^\top H \right) - \frac{1}{\mathbf{y}^\top M \mathbf{y}} \left(\mathbf{y}^\top \sigma - \mathbf{y}^\top H \mathbf{y} \right) M \mathbf{y} \mathbf{y}^\top M$$

for some matrix parameter M (induced by A).

- BFGS:** Set $M = H'$: $E^* = \frac{1}{\mathbf{y}^\top \sigma} \left(-H \mathbf{y} \sigma^\top - \sigma \mathbf{y}^\top H + \left(1 + \frac{\mathbf{y}^\top H \mathbf{y}}{\mathbf{y}^\top \sigma}\right) \sigma \sigma^\top \right)$. Equivalent update: $H' = \left(I - \frac{\sigma \mathbf{y}^\top}{\mathbf{y}^\top \sigma} \right) H \left(I - \frac{\mathbf{y} \sigma^\top}{\mathbf{y}^\top \sigma} \right) + \frac{\sigma \sigma^\top}{\mathbf{y}^\top \sigma}$.
- 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:** $\mathbf{x}_{t+1} = \Pi_X(\mathbf{x}_t - \gamma_t \mathbf{g}_t), \mathbf{g}_t \in \partial f(\mathbf{x}_t)$.
 - Lemma** (convex): $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\gamma_t (f(\mathbf{x}_t) - f^*) + \gamma_t^2 \|\mathbf{g}_t\|^2$.
- Norm of update rule $-\mathbf{x}^* \Rightarrow \Pi_X$ is non-expansive \Rightarrow Cosine theorem \Rightarrow Subgradient definition on $(\mathbf{x}^*, \mathbf{x}_t)$ (exists because of convexity). \square
- (convex): $\min_{1 \leq t \leq T} f(\mathbf{x}_t) - f^* \leq \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2 + \sum_{t=1}^T \gamma_t^2 \|\mathbf{g}_t\|^2}{2 \sum_{t=1}^T \gamma_t}$.
- Rearrange "descent" lemma \Rightarrow Sum and divide by $\sum_{t=1}^T \gamma_t$. \square
- (μ -SC, B -Lipschitz, $\gamma_t := \frac{2}{\mu(t+1)}$): $\min_{1 \leq t \leq T} f(\mathbf{x}_t) - f^* \leq \frac{2B^2}{\mu(T+1)}$.
- Adapt "descent" lemma with μ -SC \Rightarrow Def. of γ_t and $\|\mathbf{g}_t\| \leq B$. \square

Mirror descent

- Exploit non-Euclidean geometry of convex set X .
 - Bregman divergence:** Let $\omega : \Omega \rightarrow \mathbb{R}$ be continuously differentiable on Ω and 1-SC w.r.t. some norm $\|\cdot\|$. Then, $V_\omega(\mathbf{x}, \mathbf{y}) := \omega(\mathbf{x}) - \omega(\mathbf{y}) - \langle \nabla \omega(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$.
 - Properties:** $V_\omega(\mathbf{x}, \mathbf{y}) \geq 0$; $V_\omega(\mathbf{x}, \mathbf{y})$ is convex in \mathbf{x} ; $V_\omega(\mathbf{x}, \mathbf{y}) = 0$ iff $\mathbf{x} = \mathbf{y}$; $V_\omega(\mathbf{x}, \mathbf{y}) \geq \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|^2$; and $\nabla_{\mathbf{x}} V_\omega(\mathbf{x}, \mathbf{y}) = \nabla \omega(\mathbf{x}) - \nabla \omega(\mathbf{y})$.
 - 3-point id.:** $V_\omega(\mathbf{x}, \mathbf{z}) = V_\omega(\mathbf{x}, \mathbf{y}) + V_\omega(\mathbf{y}, \mathbf{z}) - \langle \nabla \omega(\mathbf{z}) - \nabla \omega(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$.
 - Update rule:** $\mathbf{x}_{t+1} \in \arg\min_{\mathbf{x} \in X} V_\omega(\mathbf{x}, \mathbf{x}_t) + \langle \gamma_t \mathbf{g}_t, \mathbf{x} \rangle, \mathbf{g}_t \in \partial f(\mathbf{x}_t)$. This is a generalization of subgradient descent.
 - Lemma:** $\gamma_t (f(\mathbf{x}_t) - f^*) \leq V_\omega(\mathbf{x}^*, \mathbf{x}_t) - V_\omega(\mathbf{x}^*, \mathbf{x}_{t+1}) + \frac{\gamma_t^2}{2} \|\mathbf{g}_t\|^2$.
- Rearrange update rule constrained optimality condition \Rightarrow 3PI $\Rightarrow -V_\omega(\mathbf{x}_{t+1}, \mathbf{x}_t) \leq -\frac{1}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \Rightarrow$ [Subgradient on $(\mathbf{x}^*, \mathbf{x}_t)$] $\cdot \gamma_t$ ($\pm \mathbf{x}_{t+1}$ in inner product) and bound with prev. \Rightarrow Young's inequality: $\langle \gamma_t \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \leq \frac{1}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \frac{1}{2} \|\gamma_t \mathbf{g}_t\|^2$. \square
- (Convex): $\min_{1 \leq t \leq T} f(\mathbf{x}_t) - f^* \leq \frac{V_\omega(\mathbf{x}^*, \mathbf{x}_0) + \frac{1}{2} \sum_{t=1}^T \gamma_t^2 \|\mathbf{g}_t\|_*^2}{\sum_{t=1}^T \gamma_t}$.
- Easily follows from above lemma by summing, dividing by summed γ_t , and telescoping sum. \square

Smoothing

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

Proximal algorithms

- Proximal operator:** $\text{prox}_{\mu f}(\mathbf{x}) := \arg\min_{\mathbf{y} \in \text{dom}(f)} f(\mathbf{y}) + \frac{1}{2\mu} \|\mathbf{x} - \mathbf{y}\|^2$.
 - Minimizer:** $\mathbf{x}^* = \text{prox}_{\mu f}(\mathbf{x}^*), \forall \mu$.
 - Non-expansiveness:** $\|\text{prox}_{\mu f}(\mathbf{x}) - \text{prox}_{\mu f}(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y}$.
 - Proximal point algorithm:** Apply gradient descent to Moreau-Yosida f_μ : $\mathbf{x}_{t+1} = \text{prox}_{\lambda_t f}(\mathbf{x}_t)$.
 - (Convex): $f(\mathbf{x}_{T+1}) - f^* \leq \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{2 \sum_{t=1}^T \lambda_t}$
- Subgradient optimality: $-\frac{\mathbf{x}_{t+1} - \mathbf{x}_t}{\lambda_t} \in \partial f(\mathbf{x}_{t+1}) \Rightarrow$ 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 $F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})$ with differentiable f (both are convex): $\mathbf{x}_{t+1} = \text{prox}_{\gamma_t g}(\mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t))$.

$(f \text{ is } L\text{-smooth, } \gamma_t := \frac{1}{L}): F(\mathbf{x}_{T+1}) - F^* \leq \frac{L\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{2T}$.

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

Stochastic optimization

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

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

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

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

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

Squared norm of update rule $-\mathbf{x}^* \Rightarrow$ Cosine theorem $\Rightarrow \mu$ -SC to get $\mathbb{E}[\langle \nabla f(\mathbf{x}_t, \xi_t), \mathbf{x}_t - \mathbf{x}^* \rangle] \geq \mu \cdot \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|^2] \Rightarrow$ Recursion. \square

- Adaptive method:** $\mathbf{g}_t = \nabla f(\mathbf{x}_t, \xi_t), \mathbf{m}_t = \phi_t(\mathbf{g}_1, \dots, \mathbf{g}_t), V_t = \psi_t(\mathbf{g}_1, \dots, \mathbf{g}_t), \hat{\mathbf{x}}_t = \mathbf{x}_t - \alpha_t V_t^{-1/2} \mathbf{m}_t, \mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in X} \left\{ \langle \mathbf{x} - \hat{\mathbf{x}}_t, V_t^{-1/2} (\mathbf{x} - \hat{\mathbf{x}}_t) \right\}$.
- SGD:** $\mathbf{m}_t = \mathbf{g}_t, V_t = I$.
- AdaGrad:** $\mathbf{m}_t = \mathbf{g}_t, V_t = \frac{\text{diag}(\sum_{\tau=1}^t \mathbf{g}_\tau^2)}{t}$.
- Adam:** $\mathbf{m}_t = (1 - \alpha) \sum_{\tau=1}^t \alpha^{t-\tau} \mathbf{g}_\tau, V_t = (1 - \beta) \text{diag}(\sum_{\tau=1}^t \beta^{t-\tau} \mathbf{g}_\tau^2)$. Recursively: $\mathbf{m}_t = \alpha \mathbf{m}_{t-1} + (1 - \alpha) \mathbf{g}_t, V_t = \beta V_{t-1} + (1 - \beta) \text{diag}(\mathbf{g}_t^2)$.

Variance reduction

- SGD requires more iterations due to high variance \Rightarrow Reduce variance.
- Finite-sum optimization:** $\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x})$.
- If we want to estimate $\theta = \mathbb{E}[X]$, we can also estimate θ as $\mathbb{E}[X - Y]$ if and only if $\mathbb{E}[Y] = 0$. Furthermore, $\text{Var}[X - Y] \leq \text{Var}[X]$ if Y is highly positively correlated with X . Specifically, if $\text{Cov}(X, Y) > \frac{1}{2} \text{Var}[Y]$, 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 $\mathbb{E}[\hat{\theta}_\alpha] = \alpha \mathbb{E}[X] + (1 - \alpha) \mathbb{E}[Y]$

$$\text{Var}[\hat{\theta}_\alpha] = \alpha^2 (\text{Var}[X] + \text{Var}[Y] - 2 \cdot \text{Cov}(X, Y)).$$

Implication: Trade-off between bias and variance, where $\alpha = 1$ makes the estimator unbiased, but the variance decreases when α decreases.

- SGD estimates $\nabla F(\mathbf{x}_t)$ by $\nabla f_{i_t}(\mathbf{x}_t)$, but VR estimates the full gradient by $\mathbf{g}_t := \alpha(\nabla f_{i_t}(\mathbf{x}_t) - Y) + \mathbb{E}[Y]$, such that \mathbf{g}_t satisfies the **VR property**: $\lim_{t \rightarrow \infty} \mathbb{E}[\|\mathbf{g}_t - \nabla F(\mathbf{x}_t)\|^2] = 0$.
- Key idea:** If \mathbf{x}_t is not too far away from previous iterates $\mathbf{x}_{1:t-1}$, we can leverage previous gradient information to construct positively correlated control variates Y .
- Stochastic Average Gradient (SAG):** Keep track of the latest gradients \mathbf{v}_i^t for all points $i \in [n]: \mathcal{O}(nd)$ storage requirement. Estimate full gradient by average of these: $\mathbf{g}_t = \frac{1}{n} \sum_{i=1}^n \mathbf{v}_i^t$. Each iteration we update \mathbf{v}_i^t by

$$\mathbf{v}_i^t = \begin{cases} \nabla f_{i_t}(\mathbf{x}_t) & i = i_t \\ \mathbf{v}_i^{t-1} & i \neq i_t. \end{cases}$$

Thus, we have $\alpha = \frac{1}{n}, Y = \mathbf{v}_i^{t-1}$, and $\mathbb{E}[Y] = \mathbf{g}_{t-1}$,

$$\mathbf{g}_t = \frac{1}{n} (\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{v}_i^{t-1}) + \mathbf{g}_{t-1}.$$

Problem: (1) $\mathcal{O}(nd)$ storage, (2) biased $\alpha \neq 1$. 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 $\alpha = 1: \mathbf{g}_t = \nabla f_{i_t}(\mathbf{x}_t) - \mathbf{v}_i^{t-1} + \mathbf{g}_{t-1}$. But, it still enjoys the same benefits.
- Stochastic variance reduced gradient (SVRG):** Build covariates based on a fixed reference point $\bar{\mathbf{x}}$ that is periodically updated every m -th iteration: $\mathbf{g}_t = \nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\bar{\mathbf{x}}) + \nabla F(\bar{\mathbf{x}})$. 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:** $\min_{\mathbf{x} \in X} \max_{\mathbf{y} \in Y} \phi(\mathbf{x}, \mathbf{y})$.
- Saddle point:** $(\mathbf{x}^*, \mathbf{y}^*)$ is a saddle point if $\phi(\mathbf{x}^*, \mathbf{y}) \leq \phi(\mathbf{x}^*, \mathbf{y}^*) \leq \phi(\mathbf{x}, \mathbf{y}^*), \forall \mathbf{x} \in X, \mathbf{y} \in Y$. Interpretation: No player has the incentive to make a unilateral change, because it can only get worse. Game theory: Nash equilibrium.
- Global minimax point:** $(\mathbf{x}^*, \mathbf{y}^*)$ is a global minimax point if $\phi(\mathbf{x}^*, \mathbf{y}) \leq \phi(\mathbf{x}^*, \mathbf{y}^*) \leq \max_{\mathbf{y}' \in Y} \phi(\mathbf{x}, \mathbf{y}'), \forall \mathbf{x} \in X, \mathbf{y} \in Y$. Interpretation: \mathbf{x}^* is the best response to the best response. Game theory: Stackelberg equilibrium.
- $\max_{\mathbf{y} \in Y} \min_{\mathbf{x} \in X} \phi(\mathbf{x}, \mathbf{y}) \leq \min_{\mathbf{x} \in X} \max_{\mathbf{y} \in Y} \phi(\mathbf{x}, \mathbf{y})$.
- Saddle point lemma:** $(\mathbf{x}^*, \mathbf{y}^*)$ is a saddle point iff $\max_{\mathbf{y} \in Y} \min_{\mathbf{x} \in X} \phi(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{x} \in X} \max_{\mathbf{y} \in Y} \phi(\mathbf{x}, \mathbf{y})$ and $(\mathbf{x}^*, \mathbf{y}^*)$ 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:** $\hat{\epsilon}(\mathbf{x}, \mathbf{y}) := \max_{\mathbf{y}' \in Y} \phi(\mathbf{x}, \mathbf{y}') - \min_{\mathbf{x}' \in X} \phi(\mathbf{x}', \mathbf{y}) \geq 0$.
- Saddle point by duality gap:** If $\hat{\epsilon}(\mathbf{x}, \mathbf{y}) = 0$, then (\mathbf{x}, \mathbf{y}) is a saddle point and if $\hat{\epsilon}(\mathbf{x}, \mathbf{y}) \leq \epsilon$, then (\mathbf{x}, \mathbf{y}) is an ϵ -saddle point.
- Gradient descent ascent (GDA):** $\mathbf{x}_{t+1} = \Pi_X(\mathbf{x}_t - \gamma \nabla_{\mathbf{x}} \phi(\mathbf{x}_t, \mathbf{y}_t)), \mathbf{y}_{t+1} = \Pi_Y(\mathbf{y}_t + \gamma \nabla_{\mathbf{y}} \phi(\mathbf{x}_t, \mathbf{y}_t))$. Does not guarantee convergence in C-C setting (consider $\phi(x, y) = xy$).

- (L -smooth, μ -SC-SC, $\gamma := \frac{\mu}{4L^2}$): $\|\mathbf{x}_T - \mathbf{x}^*\|^2 + \|\mathbf{y}_T - \mathbf{y}^*\|^2 \leq \left(1 - \frac{\mu^2}{4L^2}\right)^T (\|\mathbf{x}_1 - \mathbf{x}^*\|^2 + \|\mathbf{y}_1 - \mathbf{y}^*\|^2)$.

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

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

- Proximal point algorithm:** $(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) \in \arg\min_{\mathbf{x} \in X} \arg\max_{\mathbf{y} \in Y} \phi(\mathbf{x}, \mathbf{y}) + \frac{1}{2\gamma} \|\mathbf{x} - \mathbf{x}_t\|^2 - \frac{1}{2\gamma} \|\mathbf{y} - \mathbf{y}_t\|^2$.

Variational inequalities

- Generalizes all of the above to mapping $F: Z \rightarrow \mathbb{R}^d$. Goal: Find $\mathbf{z}^* \in Z$, such that $\langle F(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle \geq 0, \forall \mathbf{z} \in Z$.
- Monotone operator:** $\langle F(\mathbf{x}) - F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq 0$.
- μ -strongly monotone: $\langle F(\mathbf{x}) - F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \mu \|\mathbf{x} - \mathbf{y}\|^2$.
- VI strong solution (Stampacchia):** $\langle F(\mathbf{z}^*), \mathbf{z} - \mathbf{z}^* \rangle \geq 0, \forall \mathbf{z} \in Z$.
- VI weak solution (Minty):** $\langle F(\mathbf{z}), \mathbf{z} - \mathbf{z}^* \rangle \geq 0, \forall \mathbf{z} \in Z$.
- 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 $F = [\nabla_{\mathbf{x}} \phi, -\nabla_{\mathbf{y}} \phi]$ for a convex-concave ϕ .
- Extragradient method:** $\mathbf{z}_{t+1/2} = \Pi_Z(\mathbf{z}_t - \gamma_t F(\mathbf{z}_t))$
 $\mathbf{z}_{t+1} = \Pi_Z(\mathbf{z}_t - \gamma_t F(\mathbf{z}_{t+1/2}))$.
- (L -smooth, monotone, $\gamma := \frac{1}{\sqrt{2L}}$): $\max_{\mathbf{z} \in Z} \langle F(\mathbf{z}), \bar{\mathbf{z}} - \mathbf{z} \rangle \leq \frac{\sqrt{2LD_Z}}{T}$, where $\bar{\mathbf{z}} = \frac{1}{T} \sum_{t=1}^T \mathbf{z}_{t+1/2}$.

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