Definitions

- **Differentiable**: $f:\mathbb{R}^d \to \mathbb{R}$ is differentiable if
$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle + \frac{r(\boldsymbol{x}-\boldsymbol{y})}{\|\boldsymbol{x}-\boldsymbol{y}\|},$$
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**: $\forall \boldsymbol{x}\in\mathbb{R}^d$: $\boldsymbol{x}^\top A\boldsymbol{x} \geq 0$.
- B-**Lipschitz**: $\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \leq B\|\boldsymbol{x}-\boldsymbol{y}\| \Leftrightarrow \|\nabla f(\boldsymbol{x})\| \leq B$.
- **Convex set**: $\forall \boldsymbol{x},\boldsymbol{y}\in X, \lambda\in[0,1]$: $\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in X$.
- **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.
- 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] 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 $\Leftrightarrow$).
- 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] $\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle \geq \mu\|\boldsymbol{x}-\boldsymbol{y}\|^2$ (needs proof).
  [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

- **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**: $|\langle\boldsymbol{x},\boldsymbol{y}\rangle| \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\|$.
- **Hölder's inequality** (special case): $|\langle\boldsymbol{x},\boldsymbol{y}\rangle| \leq \|\boldsymbol{x}\|_1\|\boldsymbol{y}\|_\infty$.
- **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] $h(\boldsymbol{x}) = \alpha f(\boldsymbol{x}) + \beta g(\boldsymbol{x}) \Rightarrow \partial h(\boldsymbol{x}) = \alpha\cdot\partial f(\boldsymbol{x}) + \beta\cdot g(\boldsymbol{x})$.
  [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.
- 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-differentiable cases are useful when the update rule contains an argmin.

- $\boldsymbol{x}^\star$ is a local minimum.
- $\nabla f(\boldsymbol{x}^\star) = \boldsymbol{0}$.
- Constrained: $\nabla f(\boldsymbol{x}^\star)^\top(\boldsymbol{x}-\boldsymbol{x}^\star) \geq 0, \forall\boldsymbol{x}\in X$.
- 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 fundamental theorem of calculus,
$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = \int_0^1 \nabla f(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x}))^\top(\boldsymbol{y}-\boldsymbol{x})\mathrm{d}t.$$
  Or, mean-value theorem,
$$\nabla f(\boldsymbol{x}+c(\boldsymbol{y}-\boldsymbol{x}))^\top(\boldsymbol{y}-\boldsymbol{x}) = f(\boldsymbol{y}) - f(\boldsymbol{x}), \quad \exists c\in(0,1).$$
- 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: $\sum_{t=1}^T \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\| \leq \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2$.
- $f^\star \leq f(\boldsymbol{x}), \forall\boldsymbol{x}\in X$ can sometimes be useful to bound $f(\boldsymbol{x}_t) - f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - f^\star$.
- $\max\{a,b\} \leq a+b$ if $a,b\geq 0$.

Expectation and variance for SGD

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

Risk minimization

- Unknown distribution P. We only have access to samples $X_1,\ldots,X_n\sim P$. 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 $1-\delta$. (WLMM 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 $1-\delta$.

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

- **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.
  - (High $\ell_n(H_n)$, High $\ell(H_n)$): Underfitting.
  - (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**:
$$\begin{array}{ll} \text{minimize} & f_0(\boldsymbol{x}) \\ \text{subject to} & f_i(\boldsymbol{x}) \leq 0, \quad i\in[m] \\ & h_j(\boldsymbol{x}) = 0, \quad j\in[p]. \end{array}$$
- **Problem domain**: $X = \left(\bigcap_{i=0}^m \mathrm{dom}(f_i)\right) \cap \left(\bigcap_{j=1}^p \mathrm{dom}(h_j)\right)$.
- **Convex program**: All f_i are convex and all h_j are affine.

- **Lagrangian**: $L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) := f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda f_i(\boldsymbol{x}) + \sum_{j=1}^{p} \nu_j h_j(\boldsymbol{x})$.
- **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):
$$\begin{array}{ll} \text{maximize} & g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ \text{subject to} & \boldsymbol{\lambda} \geq \mathbf{0}. \end{array}$$
- 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 $\nabla_{\boldsymbol{x}} L(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) = \mathbf{0}$ (vanishing Lag. 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**: $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \gamma \nabla f(\boldsymbol{x})$.
- **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**: $f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) - \frac{1}{2L} \|\nabla f(\boldsymbol{x}_t)\|^2$.

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

    > 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 $(\boldsymbol{x}^\star, \boldsymbol{x}_T)$. $\square$

Projected gradient descent

- **Update rule** ($X \subset \mathbb{R}^d$ is closed and convex):
$$\begin{aligned} \boldsymbol{y}_{t+1} &= \boldsymbol{x}_t - \gamma_t \nabla f(\boldsymbol{x}_t) \\ \boldsymbol{x}_{t+1} &= \Pi_X(\boldsymbol{y}_{t+1}) := \operatorname*{argmin}_{\boldsymbol{x} \in X} \|\boldsymbol{x} - \boldsymbol{y}_{t+1}\|^2. \end{aligned}$$

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

   > 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

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

    > CW SD $\Rightarrow \mathbb{E}_i \Rightarrow$ Use sample prob. $\Rightarrow$ PL $\Rightarrow \mathbb{E}_{\boldsymbol{x}_t}$ (LoTE). $\square$

  - (L-smooth, μ-PL, $i \sim \mathrm{Cat}(L_1/\sum_{j=1}^d L_j, \ldots, L_d/\sum_{j=1}^d L_j)$) $\mathbb{E}[f(\boldsymbol{x}_T) - f^\star] \leq \left(1 - \frac{\mu}{dL}\right)^T (f(\boldsymbol{x}_0 - f^\star))$.

    > 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_∞, $i \in \operatorname{argmax}_{j \in [d]} |\nabla_j f(\boldsymbol{x}_t)|$) $f(\boldsymbol{x}_T) - f^\star \leq (1 - \frac{\mu}{dL})^T (f(\boldsymbol{x}_0) - f^\star)$ $f(\boldsymbol{x}_T) - f^\star \leq (1 - \frac{\mu_1}{L})^T (f(\boldsymbol{x}_0) - f^\star)$.

    > 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. $\square$

- **Trajectory analysis**: Optimize $f(\boldsymbol{x}) := \frac{1}{2}\left(\prod_{k=1}^{d} x_k - 1\right)^2$.
- $\frac{\partial f(\boldsymbol{x})}{\partial x_i} = (\prod_k x_k - 1) \prod_{k \neq i} x_k$ ($\nabla f(\boldsymbol{x}) = \mathbf{0}$ 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 $\boldsymbol{x} > 0$, $c \geq 1$. $\boldsymbol{x}$ 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.
- **Convergence** ($\gamma = \frac{1}{3dc^2}$, $\boldsymbol{x}_0 > \mathbf{0}$ 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 $\boldsymbol{x}^\star = \mathbf{1}$.

Frank-Wolfe

- $\mathrm{LMO}_X(\boldsymbol{g}) := \operatorname{argmin}_{\boldsymbol{z} \in X} \boldsymbol{g}^\top \boldsymbol{z}$.
- **Update rule**:
$$\begin{aligned} \boldsymbol{s}_t &= \mathrm{LMO}_X(\nabla f(\boldsymbol{x}_t)) \\ \boldsymbol{x}_{t+1} &= (1 - \gamma_t) \boldsymbol{x}_t + \gamma_t \boldsymbol{s}_t. \end{aligned}$$
- If $X = \mathrm{conv}(\mathcal{A})$, then $\mathrm{LMO}_X(\boldsymbol{g}) \in \mathcal{A}$.
- Advantages: (1) Iterates are always feasible if X is convex, (2) No projections, (3) Iterates have simple sparse representations as convex combination of $\{\boldsymbol{x}_0, \boldsymbol{s}_0, \ldots, \boldsymbol{s}_t\}$.
- LMO of unit ℓ_1-ball: $\mathrm{LMO}(\boldsymbol{g}) = -\mathrm{sgn}(g_i) \boldsymbol{e}_i, i \in \operatorname{argmax}_{j \in [d]} |g_i|$.
- **Optimality gap**: $g(\boldsymbol{x}) := \nabla f(\boldsymbol{x})^\top (\boldsymbol{x} - \boldsymbol{s}), \boldsymbol{s} = \mathrm{LMO}_X(\nabla f(\boldsymbol{x}))$.
- (Convex): $g(\boldsymbol{x}) \geq f(\boldsymbol{x}) - f^\star$.

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

- **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, $\gamma_t = \frac{2}{t+2}$): $f(\boldsymbol{x}_T) - f^\star \leq \frac{2L}{T+1} \mathrm{diam}(X)^2$.

> 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,
$$\begin{aligned} \nabla f'(\boldsymbol{x}') &= A^\top \nabla f(\boldsymbol{x}), \quad \boldsymbol{x} = A^{-1}(\boldsymbol{x} - \boldsymbol{b}) \\ \mathrm{LMO}_{X'}(\nabla f'(\boldsymbol{x}')) &= A^{-1}(\boldsymbol{s} - \boldsymbol{b}), \quad \boldsymbol{s} = \mathrm{LMO}_X(\nabla f(\boldsymbol{x})). \end{aligned}$$
- **Curvature constant**:
$$C_{(f,X)} := \sup_{\substack{\boldsymbol{x}, \boldsymbol{s} \in X, \gamma \in (0,1] \\ \boldsymbol{y} = (1-\gamma)\boldsymbol{x} + \gamma \boldsymbol{s}}} \frac{1}{\gamma^2} \Big(f(\boldsymbol{y}) - f(\boldsymbol{x}) - \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x})\Big).$$
- **Affine invariant convergence**: $f(\boldsymbol{x}_T) - f^\star \leq \frac{4C_{(f,X)}}{T+1}$.

> Descent lemma w.r.t. $C_{(f,X)}$ by setting $\boldsymbol{x} = \boldsymbol{x}_t, \boldsymbol{s} = \mathrm{LMO}_X(\nabla f(\boldsymbol{x}_t))$ in the supremum. $\square$

- **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.
- **Interp. 2**: Minimizes second-order Taylor approximation around $\boldsymbol{x}_t$:
$$\boldsymbol{x}_{t+1} \in \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}_t) + \nabla f(\boldsymbol{x}_t)^\top(\boldsymbol{x}-\boldsymbol{x}_t) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_t)^\top\nabla^2 f(\boldsymbol{x}_t)(\boldsymbol{x}-\boldsymbol{x}_t).$$
- **Convergence** ($\|\nabla^2 f(\boldsymbol{x})^{-1}\| \le \frac{1}{\mu}$, $\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \le B\|\boldsymbol{x}-\boldsymbol{y}\|$): $\|\boldsymbol{x}_{t+1}-\boldsymbol{x}^\star\| \le \frac{B}{2\mu}\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^2$.

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

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

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 conv. $\Rightarrow$ Update $H_t^{-1} = H_{t-1}^{-1} + E_t$ while minimizing $\|AEA^\top\|_F^2$ for some 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:
$$\begin{aligned} &\text{minimize} && \frac{1}{2}\|AEA^\top\|_F^2 \\ &\text{subject to} && E\boldsymbol{y} = \boldsymbol{r} && \text{(secant condition)} \\ & && E^\top - E = 0. && \text{(symmetry)} \end{aligned}$$
- **Greenstadt method** ($\mathcal{O}(d^2)$): Solving 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 M.
- **BFGS**: Set $M = H'$,
$$E^\star = \frac{1}{\boldsymbol{y}^\top\boldsymbol{\sigma}}\left(-H\boldsymbol{y}\boldsymbol{\sigma}^\top - \boldsymbol{\sigma}\boldsymbol{y}^\top H + \left(1 + \frac{\boldsymbol{y}^\top H\boldsymbol{y}}{\boldsymbol{y}^\top\boldsymbol{\sigma}}\right)\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\right).$$
Equivalent update:
$$H' = \left(I - \frac{\boldsymbol{\sigma}\boldsymbol{y}^\top}{\boldsymbol{y}^\top\boldsymbol{\sigma}}\right)H\left(I - \frac{\boldsymbol{y}\boldsymbol{\sigma}^\top}{\boldsymbol{y}^\top\boldsymbol{\sigma}}\right) + \frac{\boldsymbol{\sigma}\boldsymbol{\sigma}^\top}{\boldsymbol{y}^\top\boldsymbol{\sigma}}.$$
- **L-BFGS** ($\mathcal{O}(md)$): Recursive BFGS and only go down m steps.

Subgradient method

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

Norm of update rule$-\boldsymbol{x}^\star \Rightarrow \Pi_X$ is non-expansive $\Rightarrow$ Cosine theorem $\Rightarrow$ Subgradient definition (exists because of convexity). □

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

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

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

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

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,
$$V_\omega(\boldsymbol{x},\boldsymbol{y}) := \omega(\boldsymbol{x}) - \omega(\boldsymbol{y}) - \nabla\omega(\boldsymbol{y})^\top(\boldsymbol{x}-\boldsymbol{y}).$$
- **Properties**: $V_\omega(\boldsymbol{x},\boldsymbol{y}) \ge 0$; $V_\omega(\boldsymbol{x},\boldsymbol{y})$ is convex in $\boldsymbol{x}$; $V_\omega(\boldsymbol{x},\boldsymbol{y}) = 0$ iff $\boldsymbol{x} = \boldsymbol{y}$; and $V_\omega(\boldsymbol{x},\boldsymbol{y}) \ge \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2$.
- **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**: $\boldsymbol{x}_{t+1} \in \operatorname{argmin}_{\boldsymbol{x}\in X} V_\omega(\boldsymbol{x},\boldsymbol{x}_t) + \langle\gamma_t\boldsymbol{g}_t, \boldsymbol{x}\rangle, \boldsymbol{g}_t \in \partial f(\boldsymbol{x}_t)$. This is a generalization of subgradient descent.
- **Lemma**: $\gamma_t(f(\boldsymbol{x}_t) - f^\star) \le V_\omega(\boldsymbol{x}^\star,\boldsymbol{x}_t) - V_\omega(\boldsymbol{x}^\star,\boldsymbol{x}_{t+1}) + \frac{\gamma_t^2}{2}\|\gamma_t\|_\star^2$.

Rearrange update rule 1st-order optimality $\Rightarrow$ 3PI $\Rightarrow -V_\omega(\boldsymbol{x}_{t+1},\boldsymbol{x}_t) \le -\frac{1}{2}\|\boldsymbol{x}_t-\boldsymbol{x}_{t+1}\|^2 \Rightarrow$ Subgradient definition and bound with prev. $\Rightarrow$ Young's inequality: $\langle\gamma_t\boldsymbol{g}_t, \boldsymbol{x}_t-\boldsymbol{x}_{t+1}\rangle \le \frac{1}{2}\|\boldsymbol{x}_t-\boldsymbol{x}_{t+1}\|^2 + \frac{1}{2}\|\gamma_t\boldsymbol{g}_t\|_\star^2$. □

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

Easily follows from above lemma by summing, dividing by summed γ_t, and telescoping sum. □

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.
- f_μ is $1/\mu$-smooth and approximates f by $f(\boldsymbol{x}) - \mu D^2 \le f_\mu(\boldsymbol{x}) \le f(\boldsymbol{x})$, $D^2 := \max_{\boldsymbol{y}\in\mathrm{dom}(f^\star)} d(\boldsymbol{y})$.
- 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: $\min f(\boldsymbol{x}) = \min f_\mu(\boldsymbol{x})$.
- $\nabla f_\mu(\boldsymbol{x}) = \frac{1}{\mu}(\boldsymbol{x} - \mathrm{prox}_{\mu f}(\boldsymbol{x}))$.

Proximal algorithms

- **Proximal point algorithm**: Apply gradient descent to Moreau-Yosida f_μ: $\boldsymbol{x}_{t+1} = \mathrm{prox}_{\lambda_t f}(\boldsymbol{x}_t)$.
- (Convex): $f(\boldsymbol{x}_{T+1}) - f^\star \le \frac{\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{2\sum_{t=1}^T\lambda_t}$

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. □

- **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**: $\mathbb{E}_{\boldsymbol{\xi}}[\nabla f(\boldsymbol{x},\boldsymbol{\xi}) \mid \boldsymbol{x}] = \nabla F(\boldsymbol{x})$.
- **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**: $\mathbb{E}\|\nabla f(\boldsymbol{x}_t,\boldsymbol{\xi}_t) - \nabla F(\boldsymbol{x})\|^2 \le \sigma^2$.
- (L-smooth, bounded variance, random output, $\gamma := \min\left\{\frac{1}{L}, \frac{\gamma_0}{\sigma\sqrt{T}}\right\}$): $\mathbb{E}\|\nabla F(\hat{\boldsymbol{x}}_T)\|^2 \le \frac{\sigma}{\sqrt{T}}\left(\frac{2(F(\boldsymbol{x}_1)-F^\star)}{\gamma_0} + L\gamma_0\right) + \frac{2L(F(\boldsymbol{x}_1)-F^\star)}{T}$, 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: $\boldsymbol{x}_{t+1} - \boldsymbol{x}_t = -\gamma_t\nabla f(\boldsymbol{x}_t,\boldsymbol{\xi}_t) \Rightarrow \mathbb{E}[X^2] + \mathbb{E}[X]^2 + \mathrm{Var}[X]$: $\mathbb{E}\|\nabla f(\boldsymbol{x}_t,\boldsymbol{\xi}_t)\|^2 = \|\nabla F(\boldsymbol{x}_t)\|^2 + \mathbb{E}\|\nabla f(\boldsymbol{x}_t,\boldsymbol{\xi}_t) - \nabla F(\boldsymbol{x}_t)\|^2 \le \|\nabla F(\boldsymbol{x}_t)\|^2 + \sigma^2 \Rightarrow \gamma_t \le \frac{1}{L} \Rightarrow$ 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, $\mathbb{E}\|\nabla f(\boldsymbol{x},\boldsymbol{\xi})\|^2 \le B^2$) $\mathbb{E}[F(\hat{\boldsymbol{x}}_T) - F^\star] \le \frac{R^2 + B^2\sum_{t=1}^T\gamma_t^2}{2\sum_{t=1}^T\gamma_t}$, 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 expectation to bound inner product $\Rightarrow$ Convexity of $F \Rightarrow$ Telescoping sum $\Rightarrow$ Jensen's inequality. □

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

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

- **Adaptive method**: $\boldsymbol{g}_t = \nabla f(\boldsymbol{x}_t,\boldsymbol{\xi}_t)$, $\boldsymbol{m}_t = \phi_t(\boldsymbol{g}_1,\dots,\boldsymbol{g}_t)$, $V_t = \psi_t(\boldsymbol{g}_1,\dots,\boldsymbol{g}_t)$, $\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t - \alpha_t V_t^{-1/2}\boldsymbol{m}_t$, $\boldsymbol{x}_{t+1} = \operatorname{argmin}_{\boldsymbol{x}\in X}\left\{(\boldsymbol{x}-\hat{\boldsymbol{x}}_t)^\top V_t^{-1/2}(\boldsymbol{x}-\hat{\boldsymbol{x}}_t)\right\}$.
  - **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**: $\boldsymbol{m}_t = (1-\alpha)\sum_\tau^t\alpha^{t-\tau}\boldsymbol{g}_\tau$, $V_t = (1-\beta)\mathrm{diag}(\sum_{\tau=1}^t\beta^{t-\tau}\boldsymbol{g}_\tau^2)$. 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 methods estimate 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**: $\lim_{t\to\infty} \mathbb{E}\|\boldsymbol{g}_t - \nabla F(\boldsymbol{x}_t)\|^2 = 0$.
- **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
$$\boldsymbol{v}_i^t = \begin{cases} \nabla f_{i_t}(\boldsymbol{x}_t) & i = i_t \\ \boldsymbol{v}_i^{t-1} & i \neq i_t. \end{cases}$$
Thus, we have $\alpha = \frac{1}{n}$, $Y = \boldsymbol{v}_{i_t}^{t-1}$, and $\mathbb{E}[Y] = \boldsymbol{g}_{t-1}$,
$$\boldsymbol{g}_t = \frac{1}{n}\left(\nabla f_{i_t}(\boldsymbol{x}_t) - \boldsymbol{v}_{i_t}^{t-1}\right) + \boldsymbol{g}_{t-1}.$$
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 $\alpha = 1$:
$$\boldsymbol{g}_t = \nabla f_{i_t}(\boldsymbol{x}_t) - \boldsymbol{v}_{i_t}^{t-1} + \boldsymbol{g}_{t-1}.$$
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:
$$\boldsymbol{g}_t = \nabla f_{i_t}(\boldsymbol{x}_t) - \nabla f_{i_t}(\tilde{\boldsymbol{x}}) + \nabla F(\tilde{\boldsymbol{x}}).$$
Problem: (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
$$\phi(\boldsymbol{x}^\star,\boldsymbol{y}) \leq \phi(\boldsymbol{x}^\star,\boldsymbol{y}^\star) \leq \phi(\boldsymbol{x},\boldsymbol{y}^\star), \quad \forall \boldsymbol{x}\in X, \boldsymbol{y}\in Y.$$
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
$$\phi(\boldsymbol{x}^\star,\boldsymbol{y}) \leq \phi(\boldsymbol{x}^\star,\boldsymbol{y}^\star) \leq \max_{\boldsymbol{y}'\in Y}\phi(\boldsymbol{x},\boldsymbol{y}'), \quad \forall \boldsymbol{x}\in X, \boldsymbol{y}\in Y.$$
Interpretation: $\boldsymbol{x}^\star$ is the best response to the best response. Game theory: Stackelberg equilibrium.
- $\max_{\boldsymbol{y}\in Y}\min_{\boldsymbol{x}\in X}\phi(\boldsymbol{x},\boldsymbol{y}) \leq \min_{\boldsymbol{x}\in X}\max_{\boldsymbol{y}\in Y}\phi(\boldsymbol{x},\boldsymbol{y})$.
- **Saddle point lemma**: $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ is a saddle point if and only if $\max_{\boldsymbol{y}\in Y}\min_{\boldsymbol{x}\in X}\phi(\boldsymbol{x},\boldsymbol{y}) = \min_{\boldsymbol{x}\in X}\max_{\boldsymbol{y}\in Y}\phi(\boldsymbol{x},\boldsymbol{y})$.
- **Minimax theorem**: If X and Y are closed convex sets, one of them is bounded, and ϕ is a continuous convex-concave function, then there exists a saddle point in $X\times Y$.
- **Duality gap**: $\hat{\epsilon}(\boldsymbol{x},\boldsymbol{y}) := \max_{\boldsymbol{y}'\in Y}\phi(\boldsymbol{x},\boldsymbol{y}') - \min_{\boldsymbol{x}'\in X}\phi(\boldsymbol{x}',\boldsymbol{y}) \geq 0$.
- If $\hat{\epsilon}(\boldsymbol{x},\boldsymbol{y}) = 0$, then $(\boldsymbol{x},\boldsymbol{y})$ is a saddle point and if $\hat{\epsilon}(\boldsymbol{x},\boldsymbol{y}) \leq \epsilon$, 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.
- (L-smooth, μ-SC-SC, $\gamma := \frac{\mu}{4L^2}$): $\|\boldsymbol{x}_T - \boldsymbol{x}^\star\|^2 + \|\boldsymbol{y}_T - \boldsymbol{y}^\star\|^2 \leq \left(1 - \frac{\mu^2}{4L^2}\right)^T(\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + \|\boldsymbol{y}_1 - \boldsymbol{y}^\star\|^2)$.

Add μ-SC-SC definitions together ⇒ Use L-smoothness for a bound ⇒ Use update rule in $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 + \|\boldsymbol{y}_{t+1} - \boldsymbol{y}^\star\|^2$ ⇒ Non-expansiveness of projection ⇒ Rearrange ⇒ Cosine theorem ⇒ 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, $\gamma \leq \frac{1}{2L}$): $\hat{\epsilon}(\bar{\boldsymbol{x}},\bar{\boldsymbol{y}}) \leq \frac{D_X^2 + D_Y^2}{2\gamma T}$, where $\bar{\boldsymbol{x}} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_{t+1/2}$ and $\bar{\boldsymbol{y}} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{y}_{t+1/2}$.
- (L-smooth, μ-SC-SC, $\gamma := \frac{1}{8L}$): $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 + \|\boldsymbol{y}_{t+1} - \boldsymbol{y}^\star\|^2 \leq (1 - \frac{\mu}{4L})(\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + \|\boldsymbol{y}_t - \boldsymbol{y}^\star\|^2)$.
- **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 $\langle F(\boldsymbol{z}^\star), \boldsymbol{z} - \boldsymbol{z}^\star\rangle \geq 0, \forall \boldsymbol{z}\in\mathcal{Z}$.
- **Monotone operator**: $\langle F(\boldsymbol{x}) - F(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \geq 0$.
- μ-**strongly monotone**: $\langle F(\boldsymbol{x}) - F(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \geq \mu\|\boldsymbol{x} - \boldsymbol{y}\|^2$.
- **VI strong solution (Stampacchia)**: $\langle F(\boldsymbol{z}^\star), \boldsymbol{z} - \boldsymbol{z}^\star\rangle \geq 0, \forall \boldsymbol{z}\in\mathcal{Z}$.
- **VI weak solution (Minty)**: $\langle F(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{z}^\star\rangle \geq 0, \forall \boldsymbol{z}\in\mathcal{Z}$.
- If F is monotone, then strong ⇒ weak. If F is continuous, then weak ⇒ 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, $\gamma := \frac{1}{\sqrt{2}L}$): $\max_{\boldsymbol{z}\in\mathcal{Z}}\langle F(\boldsymbol{z}), \bar{\boldsymbol{z}} - \boldsymbol{z}\rangle \leq \frac{\sqrt{2}LD_{\mathcal{Z}}^2}{T}$, where $\bar{\boldsymbol{z}} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{z}_{t+1/2}$.

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

EXTRA ROOM FOR MORE INFORMATION IF NEEDED. MAYBE ADD THINGS THAT WERE USEFUL FOR THE GRADED ASSIGNMENTS. MAYBE ELABORATE ON SOME PROOFS.