Cutting Plane Algorithm

Czech Technical University in Prague
V. Franc

April 5, 2022

- Cutting Plane Algorithm
- Bundle Method for Risk Minimization
- Subgradients

XEP33SML – Structured Model Learning, Summer 2022

Structured Output SVM

- Learning $h(x;\boldsymbol{w}) = \operatorname{Argmax}_{y\in\mathcal{Y}}\langle \boldsymbol{w}, \boldsymbol{\phi}(x,y)\rangle$ from examples $\mathcal{T}^m = \{(x^i,y^i)\in(\mathcal{X}\times\mathcal{Y}) \mid i=1,\ldots,m\}$ by ERM leads to

$$\boldsymbol{w}^* \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^n} R_{\mathcal{T}^m}(\boldsymbol{w}) \quad \text{where} \quad R_{\mathcal{T}^m}(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^m \ell(y^i, h(x^i;\boldsymbol{w}))$$

Structured Output SVM

- Learning $h(x; \boldsymbol{w}) = \operatorname{Argmax}_{y \in \mathcal{Y}} \langle \boldsymbol{w}, \boldsymbol{\phi}(x, y) \rangle$ from examples $\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}$ by ERM leads to

$$\boldsymbol{w}^* \in \operatorname*{Argmin}_{\boldsymbol{w} \in \mathbb{R}^n} R_{\mathcal{T}^m}(\boldsymbol{w}) \quad \text{where} \quad R_{\mathcal{T}^m}(\boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} \ell(y^i, h(x^i; \boldsymbol{w}))$$

- The SO-SVM approximates the ERM by a convex problem

$$\boldsymbol{w}^* \in \operatorname*{Argmin}_{\boldsymbol{w} \in \mathbb{R}^n} \left(\frac{\lambda}{2} \|\boldsymbol{w}\|^2 + R^{\psi}(\boldsymbol{w}) \right) \quad \text{where} \quad R^{\psi}(\boldsymbol{w}) = \frac{1}{m} \sum_{i=1}^{m} \psi(x^i, y^i, \boldsymbol{w})$$

Structured Output SVM

- Learning $h(x;\boldsymbol{w}) = \operatorname{Argmax}_{y\in\mathcal{Y}}\langle \boldsymbol{w}, \boldsymbol{\phi}(x,y)\rangle$ from examples $\mathcal{T}^m = \{(x^i,y^i)\in(\mathcal{X}\times\mathcal{Y}) \mid i=1,\ldots,m\}$ by ERM leads to

$$\boldsymbol{w}^* \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^n} R_{\mathcal{T}^m}(\boldsymbol{w}) \quad \text{where} \quad R_{\mathcal{T}^m}(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^m \ell(y^i, h(x^i;\boldsymbol{w}))$$

- The SO-SVM approximates the ERM by a convex problem

$$\boldsymbol{w}^* \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + R^\psi(\boldsymbol{w})\right) \quad \text{where} \quad R^\psi(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^m \psi(x^i, y^i, \boldsymbol{w})$$

- The surrogate loss $\psi\colon \mathcal{X}\times\mathcal{Y}\times\mathbb{R}^n \to \mathbb{R}$ is an upper bound:

$$\ell(y, h(x;\boldsymbol{w})) \le \psi(x,y,\boldsymbol{w})\,, \quad \forall (x,y,\boldsymbol{w}) \in (\mathcal{X}\times\mathcal{Y}\times\mathbb{R}^n)$$

which is convex in $\boldsymbol{w}$ for any (x,y).

SO-SVM leads to a convex QP

- The SO-SVM with margin-rescaling loss:

$$\boldsymbol{w}^* \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \underbrace{\frac{1}{m}\sum_{i=1}^{m} \max_{y\in\mathcal{Y}} \{\ell_i(y) + \langle \boldsymbol{w}, \boldsymbol{\phi}_i(y)\rangle\}}_{R^{\psi}(\boldsymbol{w})} \right)$$

SO-SVM leads to a convex QP

- The SO-SVM with margin-rescaling loss:

$$\boldsymbol{w}^* \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \underbrace{\frac{1}{m}\sum_{i=1}^{m} \max_{y\in\mathcal{Y}}\{\ell_i(y) + \langle \boldsymbol{w}, \boldsymbol{\phi}_i(y)\rangle\}}_{R^{\psi}(\boldsymbol{w})} \right)$$

- By using slack variables it can be rewritten as a Quadratic Program:

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n, \boldsymbol{\xi}\in\mathbb{R}^m} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \xi_i \right)$$

subject to

$$\xi_i \geq \ell_i(y) + \langle \boldsymbol{w}, \boldsymbol{\phi}_i(y)\rangle\,, \quad \forall i \in \{1,\ldots,m\}, \forall y \in \mathcal{Y}$$

- Note that the QP has $m|\mathcal{Y}|$ linear constaints !

Cutting Plane Algorithm

- The SO-SVM with margin-rescaling loss:

$$\boldsymbol{w}^* \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + R^{\psi}(\boldsymbol{w}) \right)$$

- Equivalent formulation: for any $\lambda > 0$ there exists $r > 0$ such that

$$\boldsymbol{w}^* \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathcal{W}} R^{\psi}(\boldsymbol{w}) \tag{1}$$

where $\mathcal{W} = \{\boldsymbol{w} \in \mathbb{R}^n \mid \|\boldsymbol{w}\| \leq r\}$ is a ball of radius r.

Cutting Plane Algorithm

- The SO-SVM with margin-rescaling loss:

$$\boldsymbol{w}^* \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + R^\psi(\boldsymbol{w}) \right)$$

- Equivalent formulation: for any $\lambda > 0$ there exists $r > 0$ such that

$$\boldsymbol{w}^* \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathcal{W}} R^\psi(\boldsymbol{w}) \qquad (1)$$

where $\mathcal{W} = \{\boldsymbol{w} \in \mathbb{R}^n \mid \|\boldsymbol{w}\| \leq r\}$ is a ball of radius r.

- CP algorithm: approximate (1) by a series of simpler problems

$$\boldsymbol{w}_{t+1} \in \operatorname*{Argmin}_{\boldsymbol{w}\in\mathcal{W}} R_t^\psi(\boldsymbol{w}) \,, \qquad t = 1, 2, \ldots$$

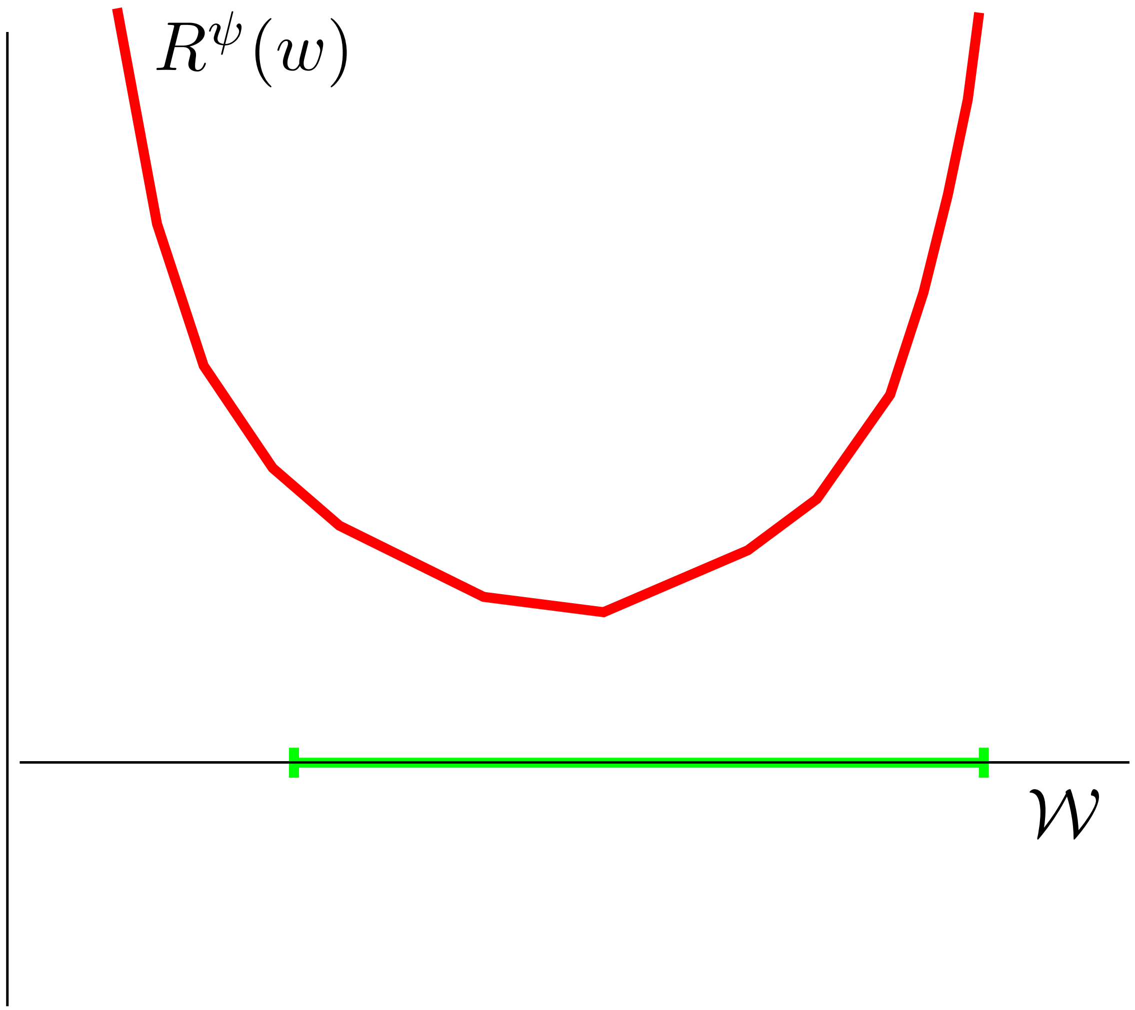
where $R_t^\psi(\boldsymbol{w})$ is a successively tighter lower bound of $R^\psi(\boldsymbol{w})$.

Cutting Plane Algorithm

$$R^{\psi}(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} \max_{\hat{y}^i \in \mathcal{Y}} (\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w}\rangle) = \max_{\substack{\hat{y}^1 \in \mathcal{Y} \\ \vdots \\ \hat{y}^m \in \mathcal{Y}}} \frac{1}{m}\sum_{i=1}^{m} (\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w}\rangle)$$

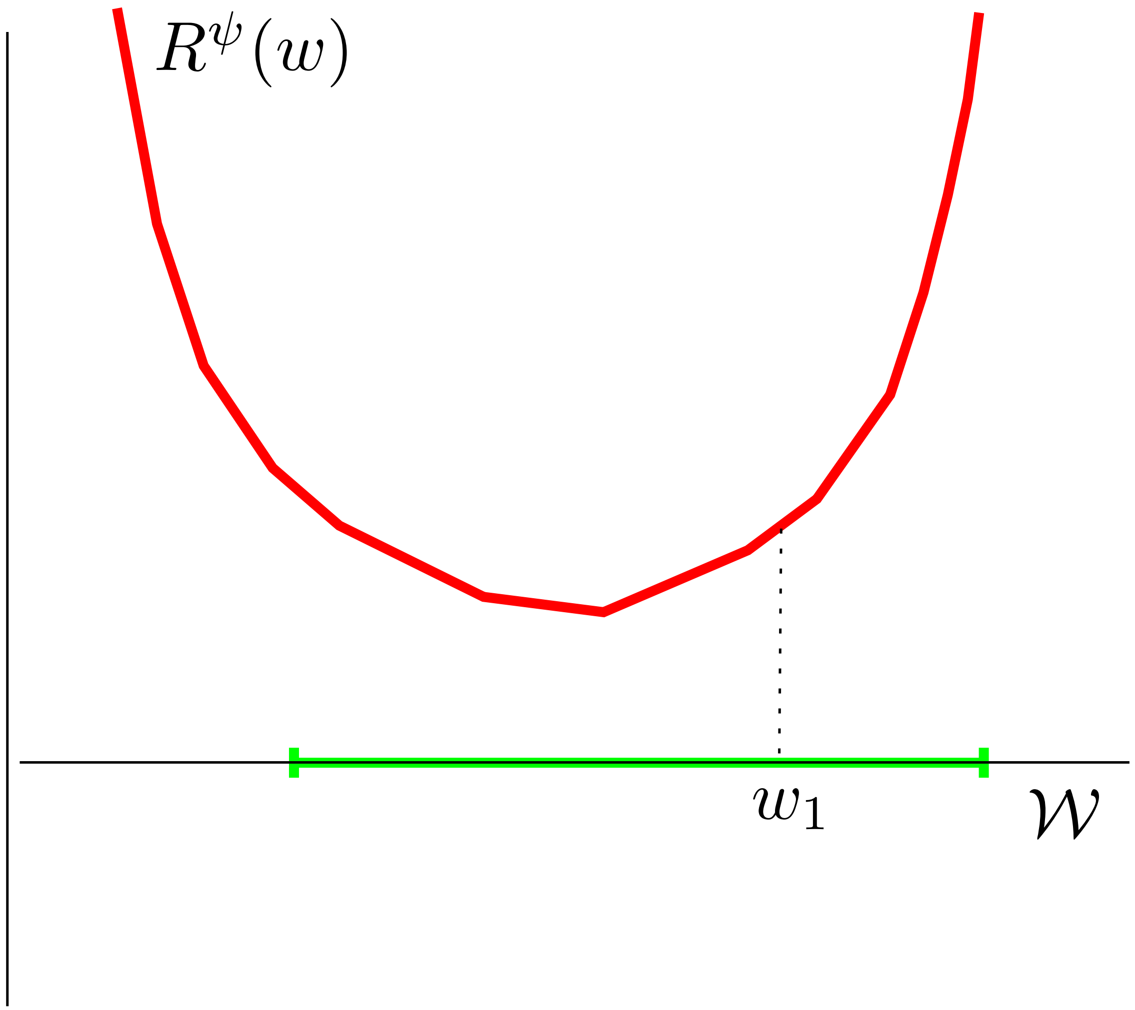
Cutting Plane Algorithm

$$R^{\psi}(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} \max_{\hat{y}^i \in \mathcal{Y}} (\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w}\rangle) = \max_{\substack{\hat{y}^1 \in \mathcal{Y} \\ \vdots \\ \hat{y}^m \in \mathcal{Y}}} \frac{1}{m}\sum_{i=1}^{m} (\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w}\rangle)$$



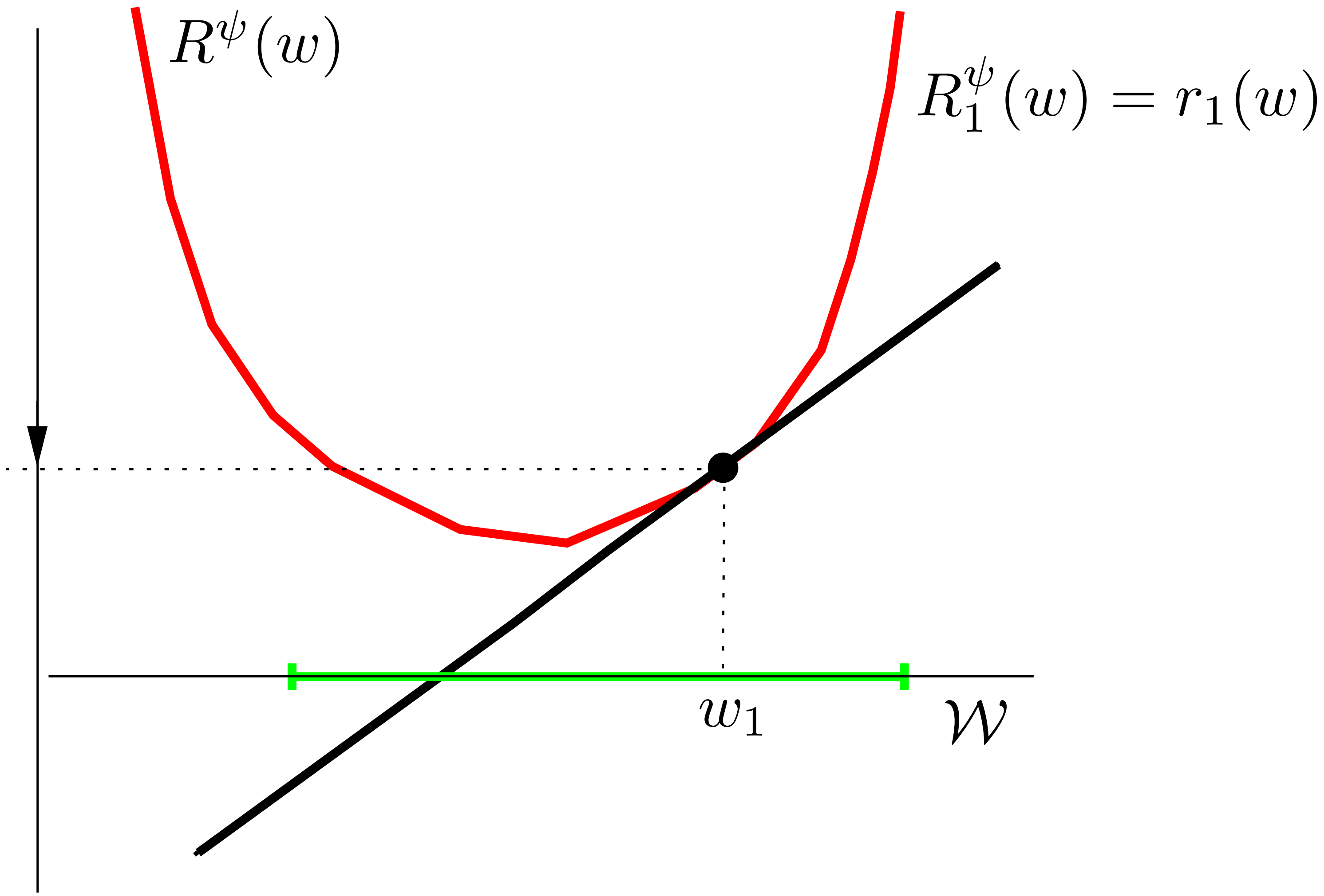
Cutting Plane Algorithm

$$R^{\psi}(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} \max_{\hat{y}^i \in \mathcal{Y}} (\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w} \rangle) = \max_{\substack{\hat{y}^1 \in \mathcal{Y} \\ \vdots \\ \hat{y}^m \in \mathcal{Y}}} \frac{1}{m}\sum_{i=1}^{m} (\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w} \rangle)$$



Cutting Plane Algorithm

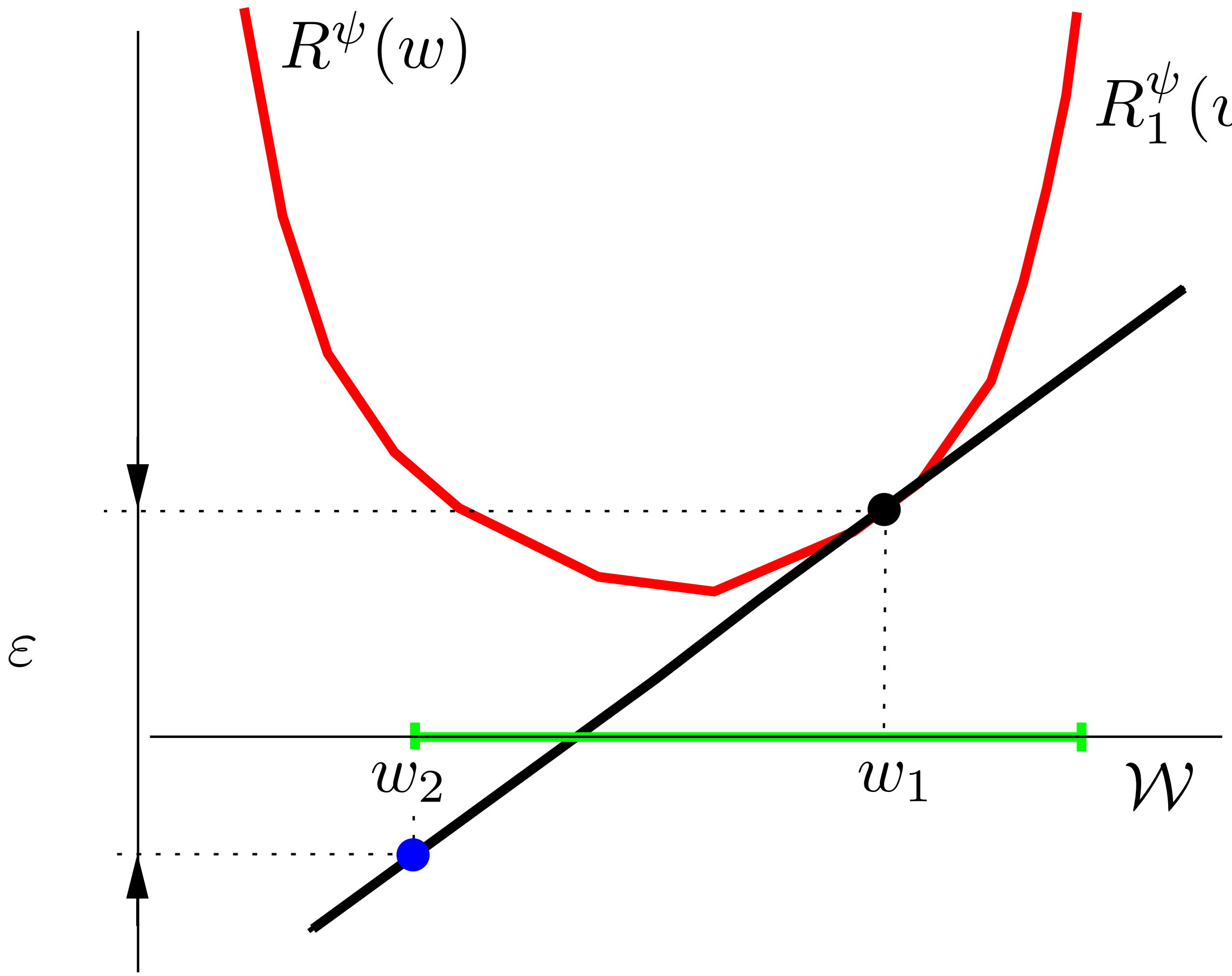
$$R^{\psi}(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} \max_{\hat{y}^i \in \mathcal{Y}} (\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w}\rangle) = \max_{\substack{\hat{y}^1 \in \mathcal{Y} \\ \vdots \\ \hat{y}^m \in \mathcal{Y}}} \frac{1}{m}\sum_{i=1}^{m} (\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w}\rangle)$$



$r_1(w) = b_1 + \langle a_1, w\rangle$

Cutting Plane Algorithm

$$R^{\psi}(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} \max_{\hat{y}^i \in \mathcal{Y}} (\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w} \rangle) = \max_{\substack{\hat{y}^1 \in \mathcal{Y} \\ \vdots \\ \hat{y}^m \in \mathcal{Y}}} \frac{1}{m}\sum_{i=1}^{m} (\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w} \rangle)$$

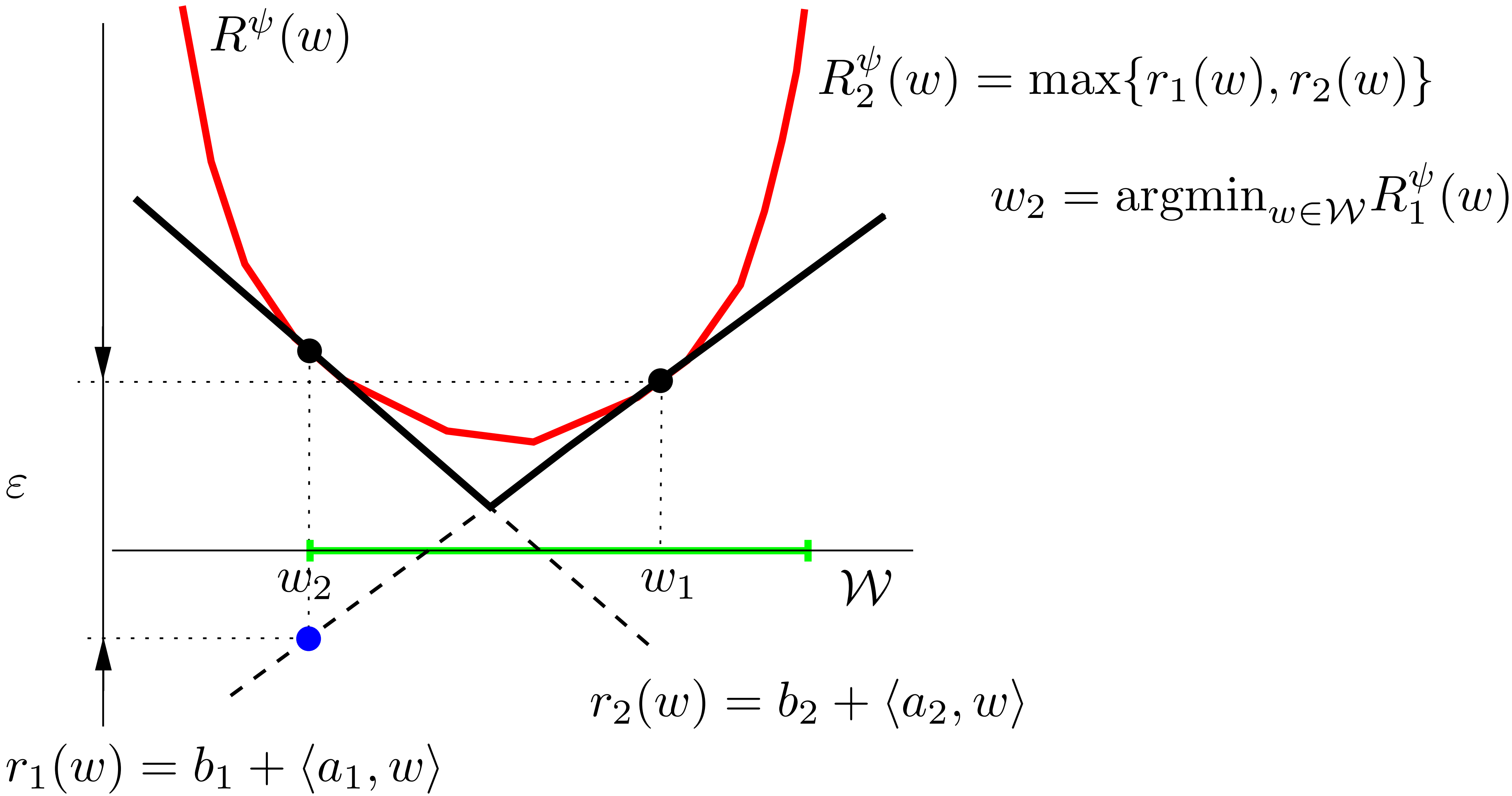


$$w_2 = \operatorname{argmin}_{w \in \mathcal{W}} R^{\psi}_1(w)$$

$$r_1(w) = b_1 + \langle a_1, w \rangle$$

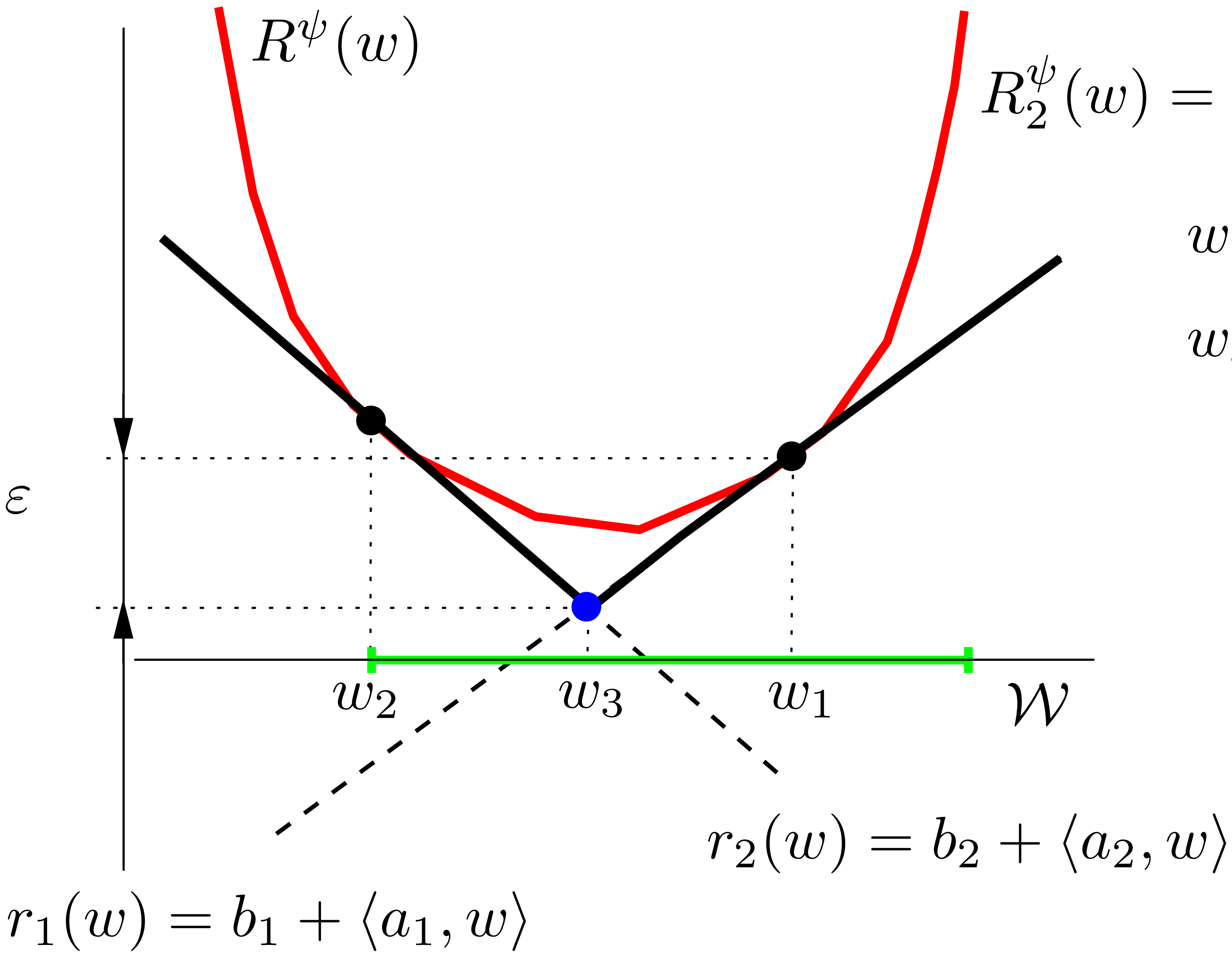
Cutting Plane Algorithm

$$R^\psi(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} \max_{\hat{y}^i\in\mathcal{Y}}(\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w}\rangle) = \max_{\substack{\hat{y}^1\in\mathcal{Y}\\ \vdots \\ \hat{y}^m\in\mathcal{Y}}} \frac{1}{m}\sum_{i=1}^{m}(\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w}\rangle)$$



Cutting Plane Algorithm

$$R^{\psi}(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} \max_{\hat{y}^i \in \mathcal{Y}} (\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w} \rangle) = \max_{\substack{\hat{y}^1 \in \mathcal{Y} \\ \vdots \\ \hat{y}^m \in \mathcal{Y}}} \frac{1}{m}\sum_{i=1}^{m} (\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w} \rangle)$$

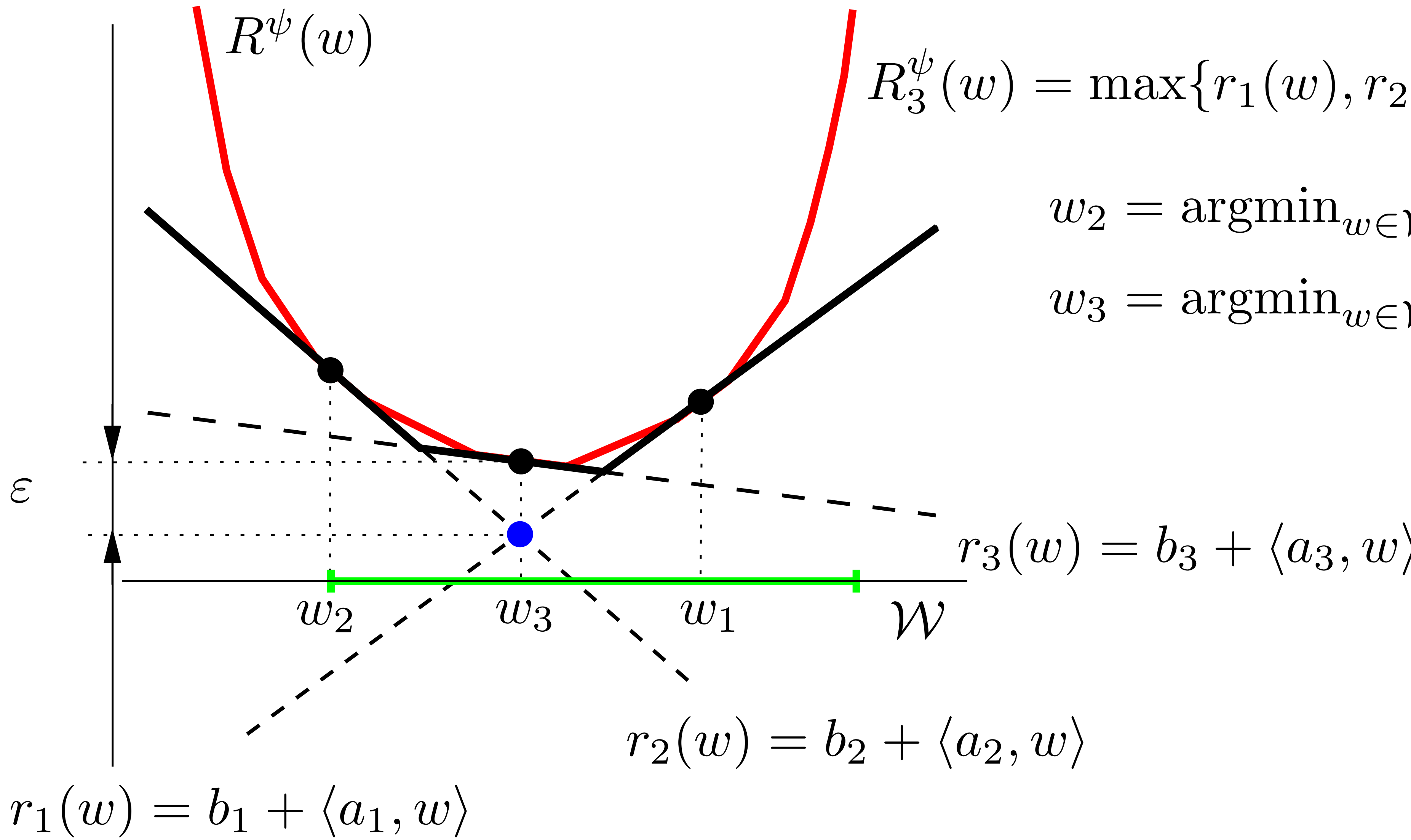


$$w_2 = \text{argmin}_{w \in \mathcal{W}} R_1^{\psi}(w)$$

$$w_3 = \text{argmin}_{w \in \mathcal{W}} R_2^{\psi}(w)$$

Cutting Plane Algorithm

$$R^{\psi}(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} \max_{\hat{y}^i \in \mathcal{Y}} (\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w}\rangle) = \max_{\substack{\hat{y}^1 \in \mathcal{Y} \\ \vdots \\ \hat{y}^m \in \mathcal{Y}}} \frac{1}{m}\sum_{i=1}^{m} (\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w}\rangle)$$

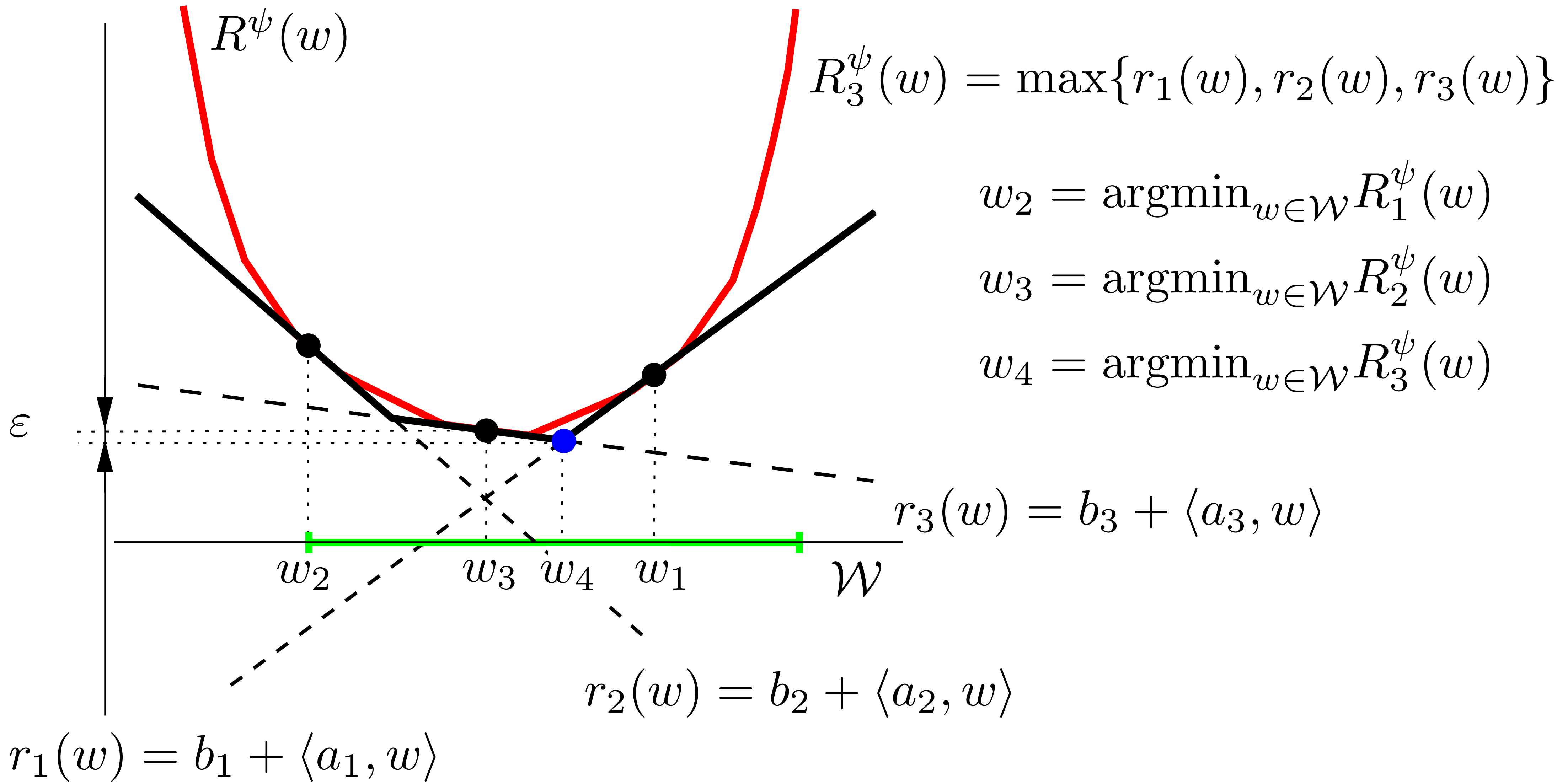


$$w_2 = \mathrm{argmin}_{w\in\mathcal{W}} R^{\psi}_1(w)$$

$$w_3 = \mathrm{argmin}_{w\in\mathcal{W}} R^{\psi}_2(w)$$

Cutting Plane Algorithm

$$R^{\psi}(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} \max_{\hat{y}^i\in\mathcal{Y}}(\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w}\rangle) = \max_{\substack{\hat{y}^1\in\mathcal{Y}\\ \vdots \\ \hat{y}^m\in\mathcal{Y}}} \frac{1}{m}\sum_{i=1}^{m}(\ell_i(\hat{y}^i) + \langle \boldsymbol{\phi}_i(\hat{y}^i), \boldsymbol{w}\rangle)$$



$$w_2 = \text{argmin}_{w\in\mathcal{W}} R_1^{\psi}(w)$$

$$w_3 = \text{argmin}_{w\in\mathcal{W}} R_2^{\psi}(w)$$

$$w_4 = \text{argmin}_{w\in\mathcal{W}} R_3^{\psi}(w)$$

Cutting Plane Algorithm

1. $\boldsymbol{w}_1 \in \mathcal{W} = \{\boldsymbol{w} \in \mathbb{R}^n \mid \|\boldsymbol{w}\| \leq r\}$, $t \leftarrow 1$

2. Compute a new cutting plane and the objective value:

$$\boldsymbol{a}_t = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\phi}_i(\hat{y}^i)\,, \quad b_t = \frac{1}{m}\sum_{i=1}^{m} \ell_i(\hat{y}^i)\,, \quad R^{\psi}(\boldsymbol{w}_t) = b_t + \langle \boldsymbol{w}_t, \boldsymbol{a}_t \rangle$$

   where $\hat{y}^i$ is a solutions of loss augmented prediction problem:

$$\hat{y}^i = \operatorname*{argmax}_{y \in \mathcal{Y}} \left(\ell_i(y) + \langle \boldsymbol{w}, \boldsymbol{\phi}_i(y) \rangle \right) = \operatorname*{argmax}_{y \in \mathcal{Y}} \left(\ell(y^i, y) + \langle \boldsymbol{w}, \boldsymbol{\phi}(x^i, y) \rangle \right)$$

3. Solve a reduced problem

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathcal{W}} R_t^{\psi}(\boldsymbol{w}) \quad \text{where} \quad R_t^{\psi}(\boldsymbol{w}) = \max_{i=1,\ldots,t} (b_i + \langle \boldsymbol{w}, \boldsymbol{a}_i \rangle)$$

4. If $\min_{i=1,\ldots,t} R^{\psi}(\boldsymbol{w}_t) - R_t^{\psi}(\boldsymbol{w}_{t+1}) \leq \varepsilon$ exit else $t \leftarrow t+1$ and go to 2.

Bundle Method for Risk Minimization

1. $\boldsymbol{w}_1 \in \mathbb{R}^n$, $t \leftarrow 1$

2. Compute a new cutting plane and the objective value:

$$\boldsymbol{a}_t = \frac{1}{m}\sum_{i=1}^m \boldsymbol{\phi}_i(\hat{y}^i)\,, \quad b_t = \frac{1}{m}\sum_{i=1}^m \ell_i(\hat{y}^i)\,, \quad R^\psi(\boldsymbol{w}_t) = b_t + \langle \boldsymbol{w}_t, \boldsymbol{a}_t\rangle$$

where $\hat{y}^i$ is a solutions of loss augmented prediction problem:

$$\hat{y}^i = \operatorname*{argmax}_{y\in\mathcal{Y}} \big(\ell_i(y) + \langle \boldsymbol{w}, \boldsymbol{\phi}_i(y)\rangle\big) = \operatorname*{argmax}_{y\in\mathcal{Y}} \big(\ell(y^i, y) + \langle \boldsymbol{w}, \boldsymbol{\phi}(x^i, y)\rangle\big)$$

3. Solve a reduced problem

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + R_t^\psi(\boldsymbol{w})\right) \quad \text{where} \quad R_t^\psi(\boldsymbol{w}) = \max_{i=1,\dots,t}(b_i + \langle \boldsymbol{w}, \boldsymbol{a}_i\rangle)$$

4. If $\min_{i=1,\dots,t}\big(\frac{\lambda}{2}\|\boldsymbol{w}_i\|^2 + R^\psi(\boldsymbol{w}_i)\big) - \big(\frac{\lambda}{2}\|\boldsymbol{w}_{t+1}\|^2 + R_t^\psi(\boldsymbol{w}_{t+1})\big) \leq \varepsilon$ exit else $t \leftarrow t+1$ and go to 2.

Bundle Method for Risk Minimization

- The original convex problem

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n} F(\boldsymbol{w}) := \left(\frac{\lambda}{2} \|\boldsymbol{w}\|^2 + R(\boldsymbol{w}) \right)$$

Bundle Method for Risk Minimization

- The original convex problem

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} F(\boldsymbol{w}) := \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + R(\boldsymbol{w})\right)$$

is reduced to a sequence of reduced convex problems

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} F_t(\boldsymbol{w}) := \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + R_t(\boldsymbol{w})\right)$$

Bundle Method for Risk Minimization

- The original convex problem

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} F(\boldsymbol{w}) := \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + R(\boldsymbol{w})\right)$$

is reduced to a sequence of reduced convex problems

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} F_t(\boldsymbol{w}) := \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + R_t(\boldsymbol{w})\right)$$

where $R_t(\boldsymbol{w})$ is the “cutting plane model”

$$R_t(\boldsymbol{w}) = \max_{i=0,\ldots,t}\left[R(\boldsymbol{w}_i) + \langle \boldsymbol{g}_i, \boldsymbol{w} - \boldsymbol{w}_i\rangle\right]$$

and $\boldsymbol{g}_i = \partial R(\boldsymbol{w}_i)$ is a subgradient of $R(\boldsymbol{w})$ at $\boldsymbol{w}_i$.

Bundle Method for Risk Minimization

- The original convex problem

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} F(\boldsymbol{w}) := \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + R(\boldsymbol{w})\right)$$

is reduced to a sequence of reduced convex problems

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} F_t(\boldsymbol{w}) := \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + R_t(\boldsymbol{w})\right)$$

where $R_t(\boldsymbol{w})$ is the "cutting plane model"

$$R_t(\boldsymbol{w}) = \max_{i=0,\ldots,t} \left[R(\boldsymbol{w}_i) + \langle \boldsymbol{g}_i, \boldsymbol{w} - \boldsymbol{w}_i\rangle\right]$$

and $\boldsymbol{g}_i = \partial R(\boldsymbol{w}_i)$ is a subgradient of $R(\boldsymbol{w})$ at $\boldsymbol{w}_i$.

- By construction it holds that $R_t(\boldsymbol{w}) \leq R(\boldsymbol{w})$, $\forall \in \mathbb{R}^n$.

Subgradient

- Let $f\colon \mathcal{X} \to \mathbb{R}$ be a convex function where $\mathcal{X} \subseteq \mathbb{R}^n$ is a convex set.

Subgradient

- Let $f\colon \mathcal{X} \to \mathbb{R}$ be a convex function where $\mathcal{X} \subseteq \mathbb{R}^n$ is a convex set.
- For differentiable f the gradient $\nabla f(\boldsymbol{x}') \in \mathbb{R}^n$ at point $\boldsymbol{x}' \in \mathcal{X}$ determines a global under-estimator of f

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}') + \nabla f(\boldsymbol{x}')^T(\boldsymbol{x} - \boldsymbol{x}') \,, \quad \forall \boldsymbol{x} \in \mathcal{X} \,.$$

Subgradient

- Let $f\colon \mathcal{X} \to \mathbb{R}$ be a convex function where $\mathcal{X} \subseteq \mathbb{R}^n$ is a convex set.
- For differentiable f the gradient $\nabla f(\boldsymbol{x}') \in \mathbb{R}^n$ at point $\boldsymbol{x}' \in \mathcal{X}$ determines a global under-estimator of f

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}') + \nabla f(\boldsymbol{x}')^T(\boldsymbol{x} - \boldsymbol{x}')\,, \quad \forall \boldsymbol{x} \in \mathcal{X}\,.$$

- For non-differentiable f we can still construct a global under-estimator: the vector $\boldsymbol{g} \in \mathbb{R}^n$ is a subgradient of f at point $\boldsymbol{x}' \in \mathcal{X}$ if

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}') + \boldsymbol{g}^T(\boldsymbol{x} - \boldsymbol{x}')\,, \quad \forall \boldsymbol{x} \in \mathcal{X}\,.$$

Subgradient

- Let $f\colon \mathcal{X} \to \mathbb{R}$ be a convex function where $\mathcal{X} \subseteq \mathbb{R}^n$ is a convex set.
- For differentiable f the gradient $\nabla f(\boldsymbol{x}') \in \mathbb{R}^n$ at point $\boldsymbol{x}' \in \mathcal{X}$ determines a global under-estimator of f

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}') + \nabla f(\boldsymbol{x}')^T(\boldsymbol{x} - \boldsymbol{x}')\,, \quad \forall \boldsymbol{x} \in \mathcal{X}\,.$$

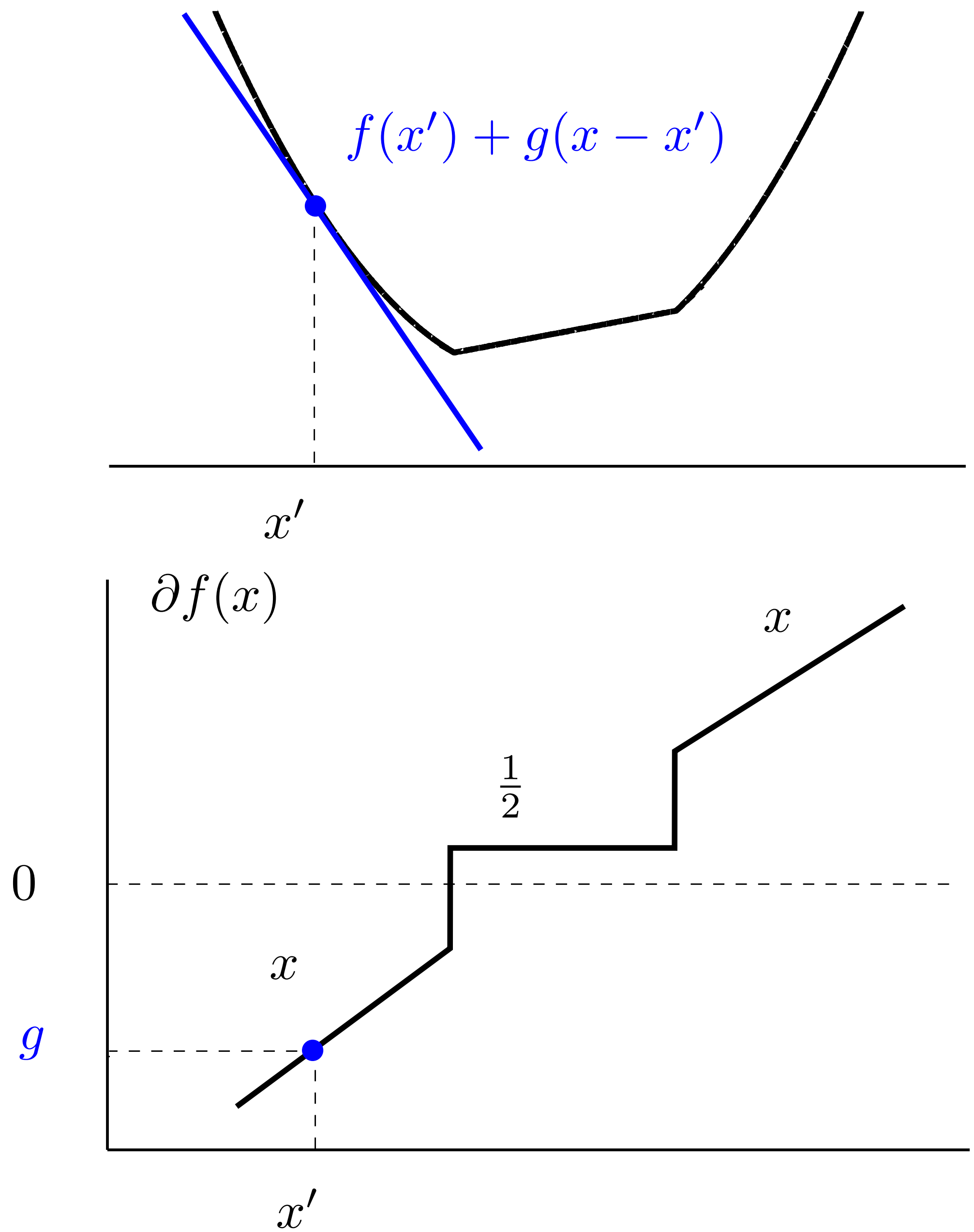
- For non-differentiable f we can still construct a global under-estimator: the vector $\boldsymbol{g} \in \mathbb{R}^n$ is a subgradient of f at point $\boldsymbol{x}' \in \mathcal{X}$ if

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}') + \boldsymbol{g}^T(\boldsymbol{x} - \boldsymbol{x}')\,, \quad \forall \boldsymbol{x} \in \mathcal{X}\,.$$

- There can be more then one subgradient at given point: the collection of subgradients of f at point $\boldsymbol{x} \in \mathcal{X}$ is the subdifferential $\partial f(\boldsymbol{x})$.

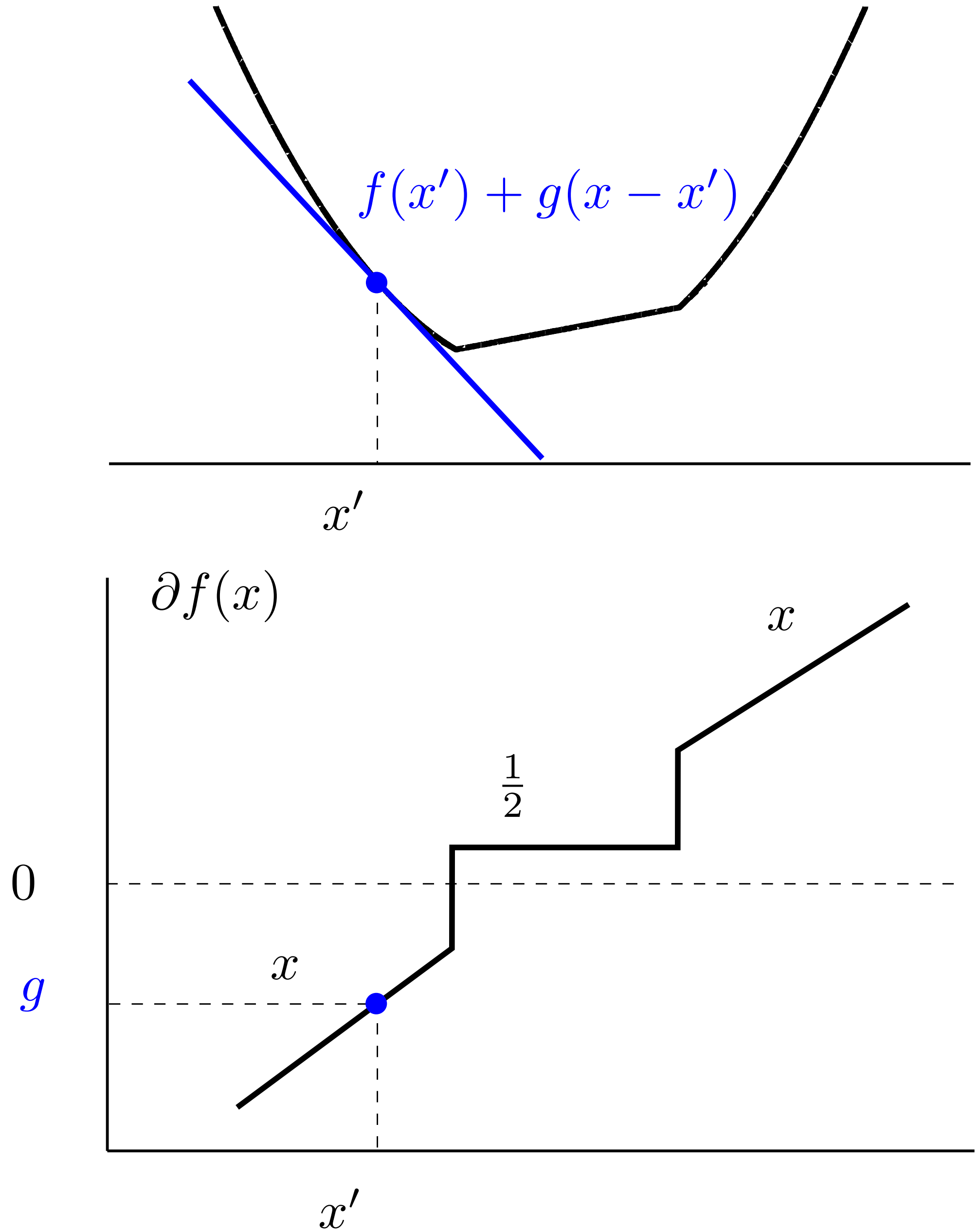
An example: non-differentiable function and its subgradients

$$f(x) = \max\{\tfrac{1}{2}x^2, \tfrac{1}{2}x + 2\}$$



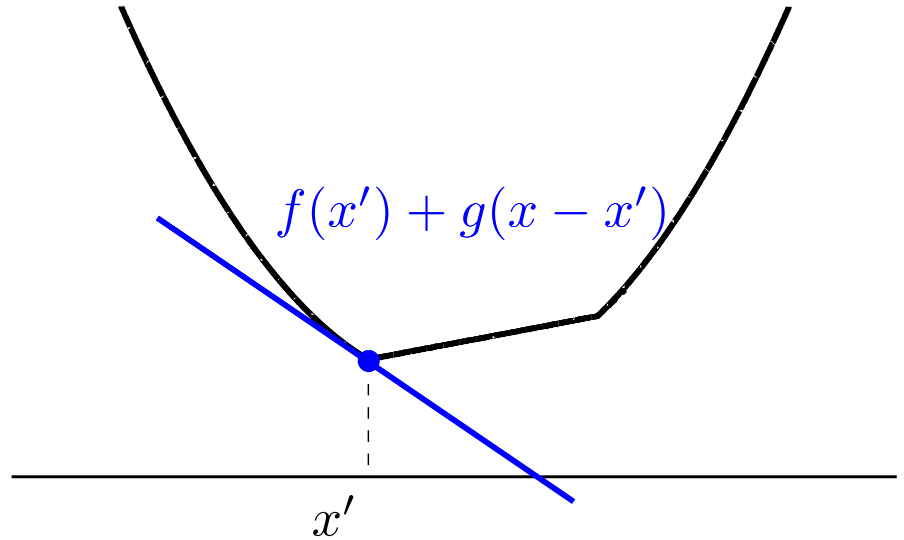
An example: non-differentiable function and its subgradients

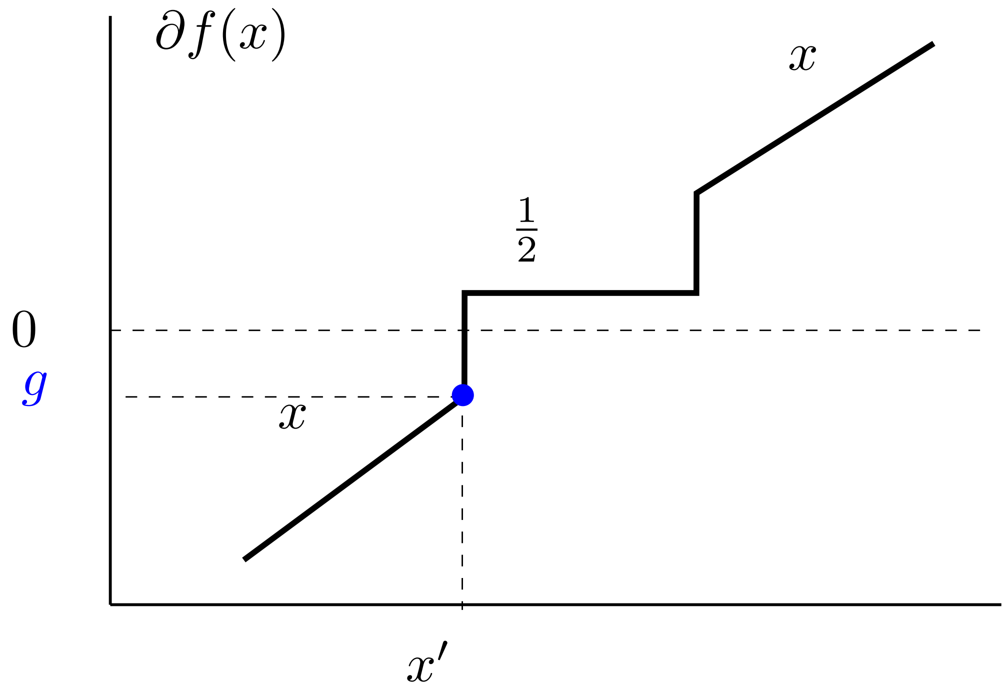
$$f(x) = \max\{\tfrac{1}{2}x^2, \tfrac{1}{2}x + 2\}$$



An example: non-differentiable function and its subgradients

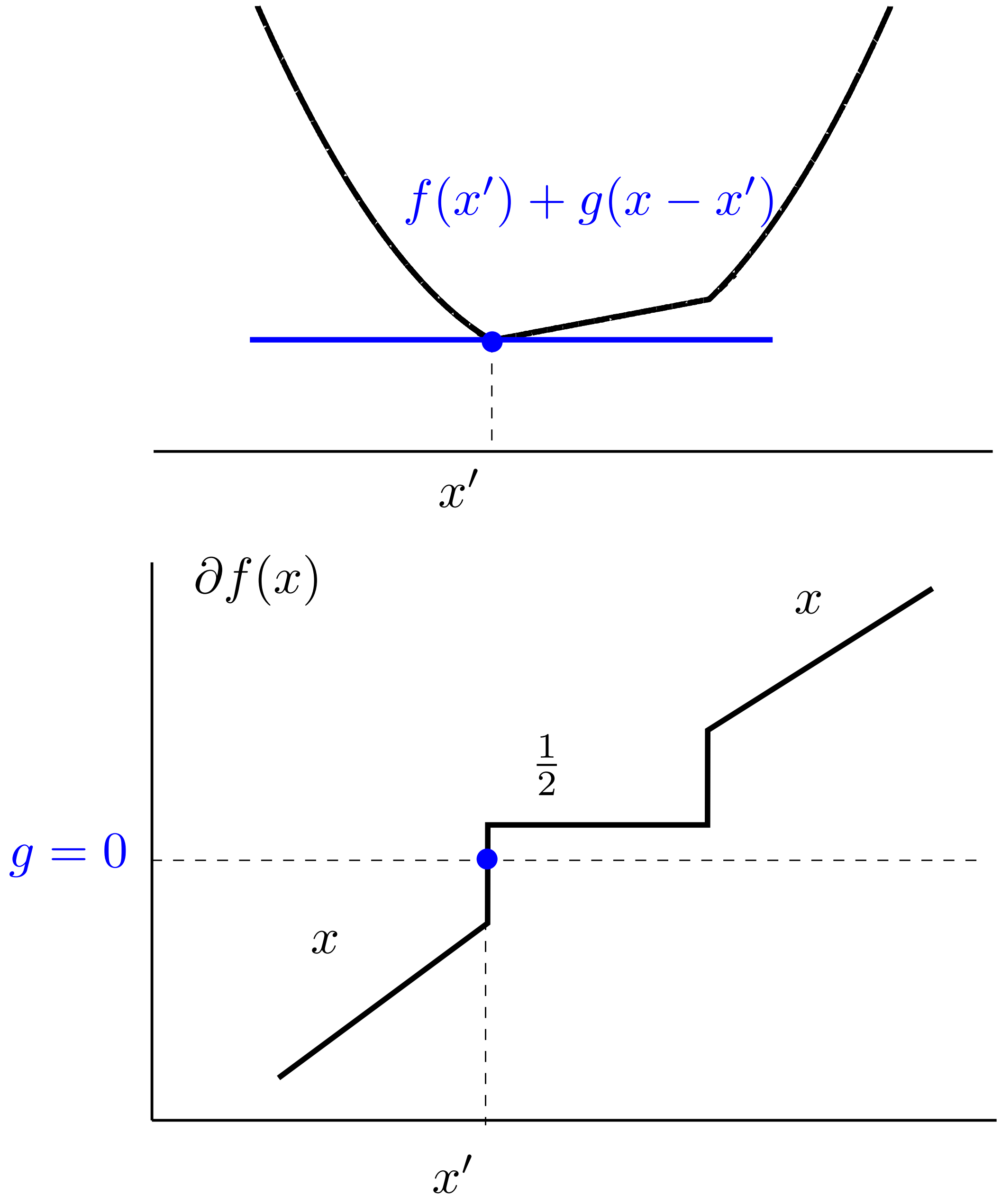
$$f(x) = \max\{\tfrac{1}{2}x^2, \tfrac{1}{2}x + 2\}$$





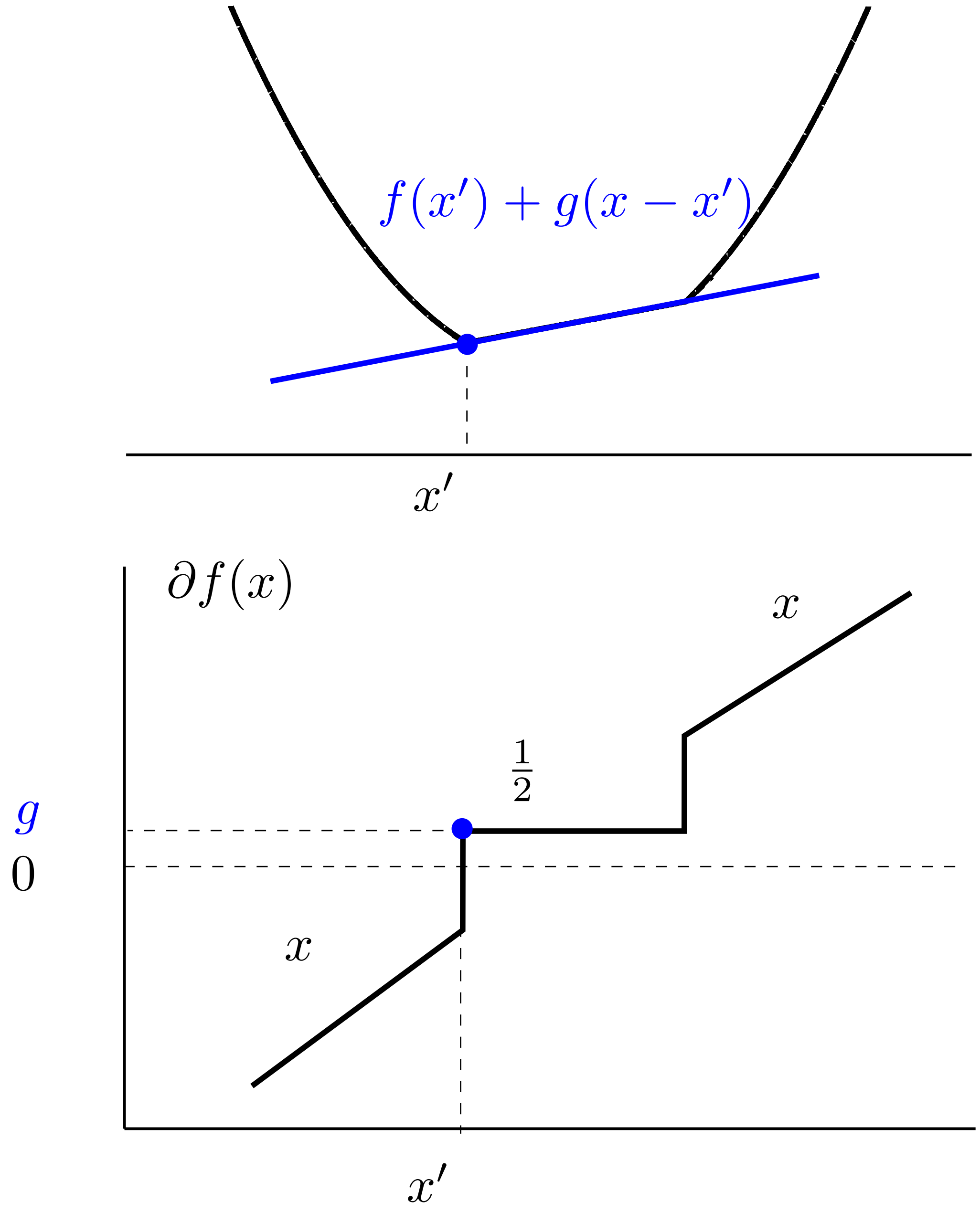
An example: non-differentiable function and its subgradients

$$f(x) = \max\{\tfrac{1}{2}x^2, \tfrac{1}{2}x + 2\}$$



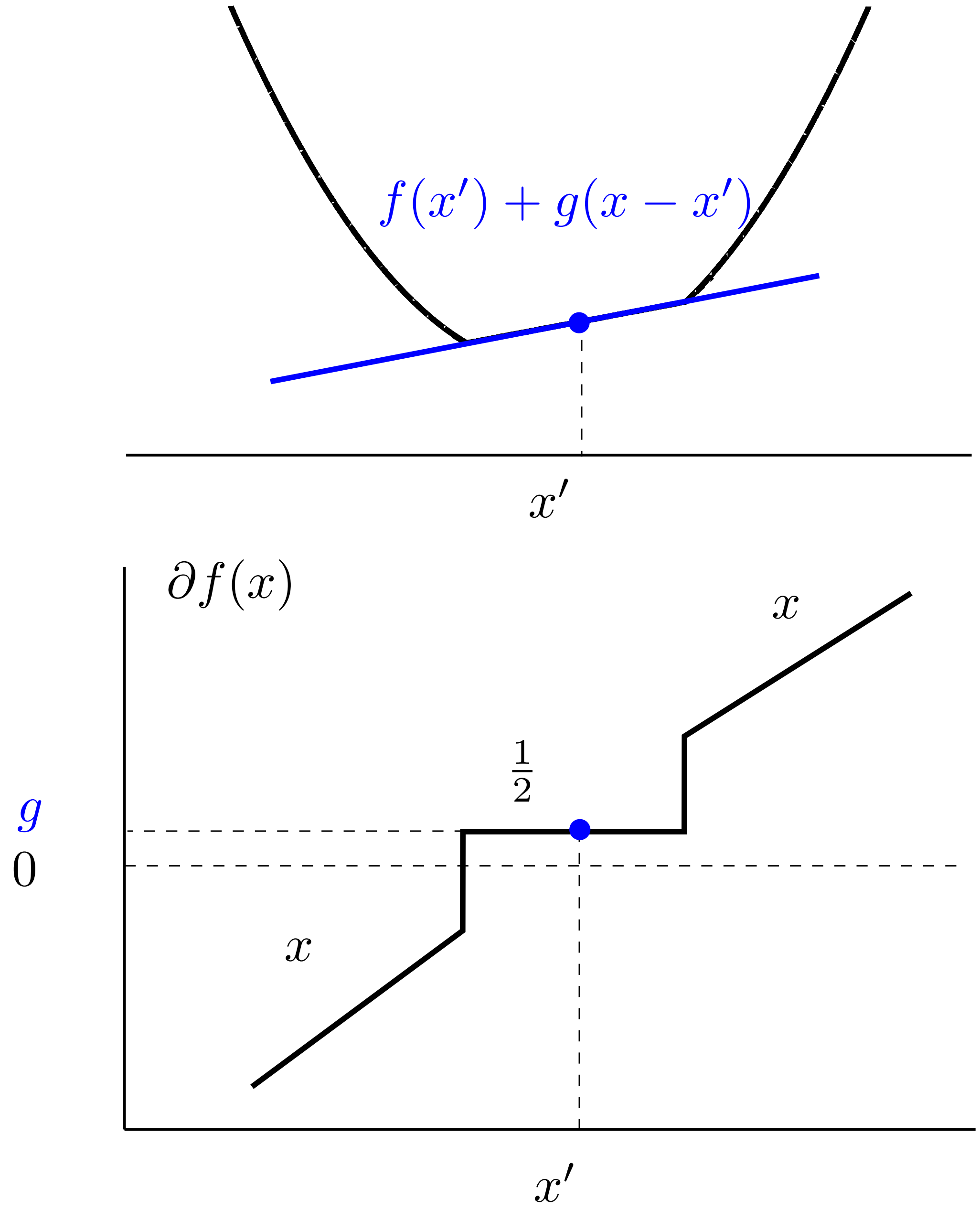
An example: non-differentiable function and its subgradients

$$f(x) = \max\{\tfrac{1}{2}x^2, \tfrac{1}{2}x + 2\}$$



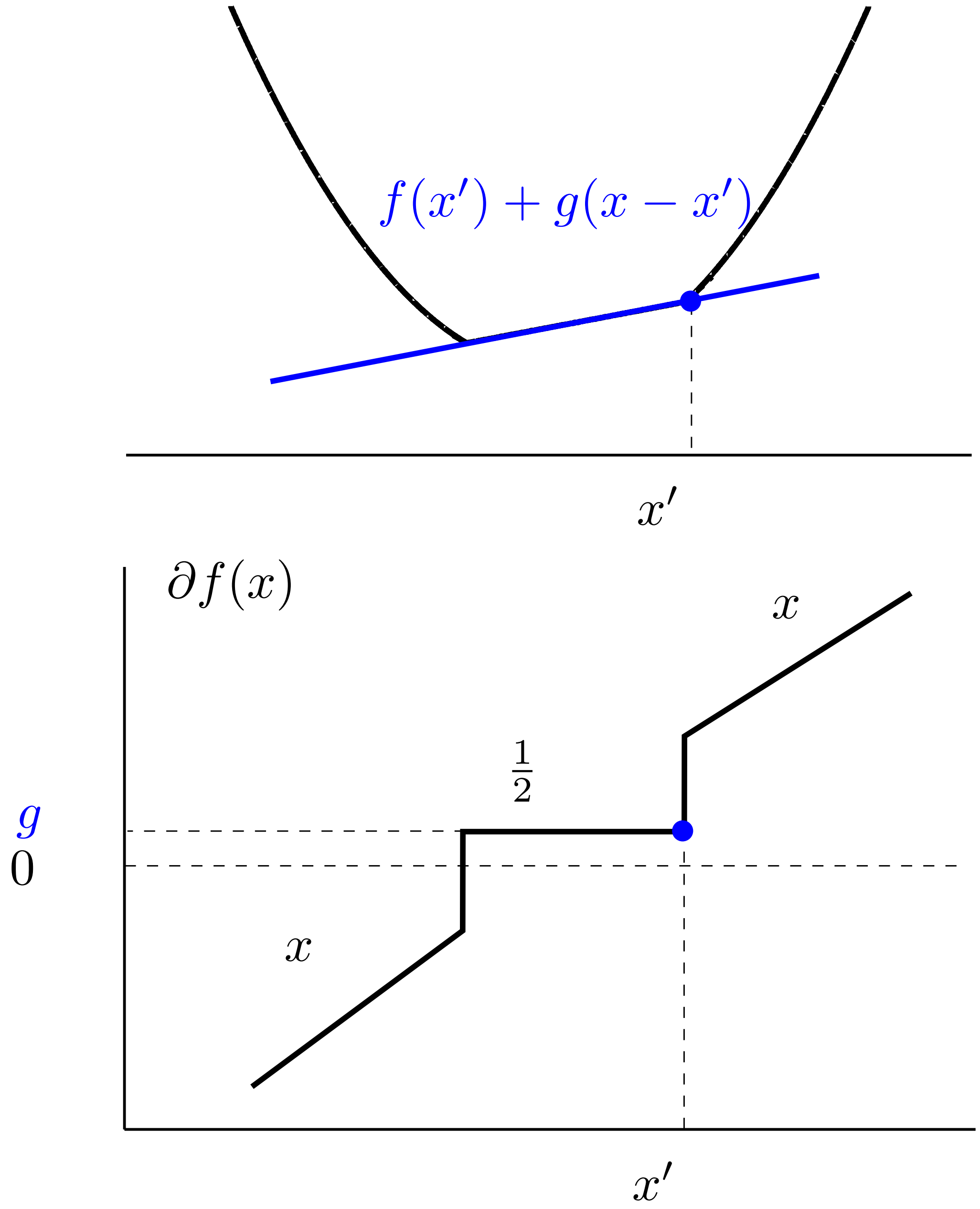
An example: non-differentiable function and its subgradients

$$f(x) = \max\{\tfrac{1}{2}x^2, \tfrac{1}{2}x + 2\}$$



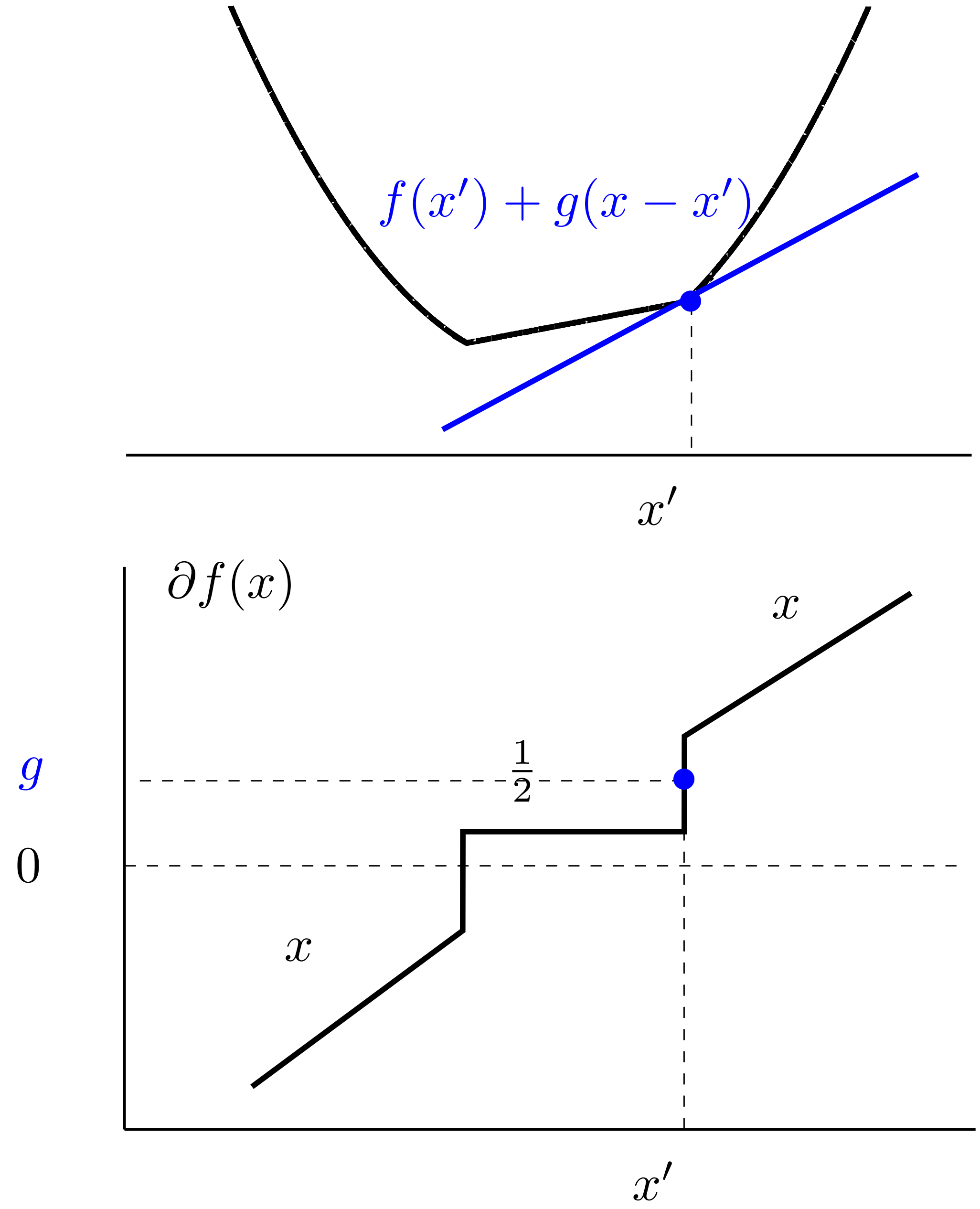
An example: non-differentiable function and its subgradients

$$f(x) = \max\{\tfrac{1}{2}x^2, \tfrac{1}{2}x + 2\}$$



An example: non-differentiable function and its subgradients

$$f(x) = \max\{\tfrac{1}{2}x^2, \tfrac{1}{2}x + 2\}$$



Basic subgradient calculus

- $\partial f(\boldsymbol{x})$ is a closed convex set.

Basic subgradient calculus

- $\partial f(\boldsymbol{x})$ is a closed convex set.
- $\partial f(\boldsymbol{x}) = \{\boldsymbol{g}\} \iff f$ is differentiable and $\boldsymbol{g} = \nabla f(\boldsymbol{x})$.

Basic subgradient calculus

- $\partial f(\boldsymbol{x})$ is a closed convex set.
- $\partial f(\boldsymbol{x}) = \{\boldsymbol{g}\} \iff f$ is differentiable and $\boldsymbol{g} = \nabla f(\boldsymbol{x})$.
- scaling: $\partial(\alpha f(\boldsymbol{x})) = \alpha \partial f(\boldsymbol{x})$ if $\alpha > 0$.

Basic subgradient calculus

- $\partial f(\boldsymbol{x})$ is a closed convex set.
- $\partial f(\boldsymbol{x}) = \{\boldsymbol{g}\} \iff f$ is differentiable and $\boldsymbol{g} = \nabla f(\boldsymbol{x})$.
- scaling: $\partial(\alpha f(\boldsymbol{x})) = \alpha \partial f(\boldsymbol{x})$ if $\alpha > 0$.
- addition: $\partial(f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})) = \partial f_1(\boldsymbol{x}) + \partial f_2(\boldsymbol{x})$.

Basic subgradient calculus

- $\partial f(\boldsymbol{x})$ is a closed convex set.
- $\partial f(\boldsymbol{x}) = \{\boldsymbol{g}\} \iff f$ is differentiable and $\boldsymbol{g} = \nabla f(\boldsymbol{x})$.
- scaling: $\partial(\alpha f(\boldsymbol{x})) = \alpha \partial f(\boldsymbol{x})$ if $\alpha > 0$.
- addition: $\partial(f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})) = \partial f_1(\boldsymbol{x}) + \partial f_2(\boldsymbol{x})$.
- point-wise maximum: $f(\boldsymbol{x}) = \max_{i=1,\ldots,m} f_i(\boldsymbol{x})$ when $f_i(\boldsymbol{x})$ are differentiable then

$$\partial f(\boldsymbol{x}) = \mathbf{Co}\{\nabla f_i(\boldsymbol{x}) \mid f_i(\boldsymbol{x}) = f(\boldsymbol{x})\} ,$$

i.e., convex hull of gradients of active functions at $\boldsymbol{x}$.

Basic subgradient calculus

- $\partial f(\boldsymbol{x})$ is a closed convex set.
- $\partial f(\boldsymbol{x}) = \{\boldsymbol{g}\} \iff f$ is differentiable and $\boldsymbol{g} = \nabla f(\boldsymbol{x})$.
- scaling: $\partial(\alpha f(\boldsymbol{x})) = \alpha \partial f(\boldsymbol{x})$ if $\alpha > 0$.
- addition: $\partial(f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})) = \partial f_1(\boldsymbol{x}) + \partial f_2(\boldsymbol{x})$.
- point-wise maximum: $f(\boldsymbol{x}) = \max_{i=1,\ldots,m} f_i(\boldsymbol{x})$ when $f_i(\boldsymbol{x})$ are differentiable then

$$\partial f(\boldsymbol{x}) = \mathbf{Co}\{\nabla f_i(\boldsymbol{x}) \mid f_i(\boldsymbol{x}) = f(\boldsymbol{x})\} ,$$

  i.e., convex hull of gradients of active functions at $\boldsymbol{x}$.
- Optimality condition for a convex f:

$$f(\boldsymbol{x}^*) = \inf_{\boldsymbol{x}} f(\boldsymbol{x}) \iff \mathbf{0} \in \partial f(\boldsymbol{x}^*)$$

Example: cutting plane model for SVM

- The convex problem to solver

$$\boldsymbol{w}^* \in \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \underbrace{\frac{1}{m}\sum_{i=1}^{m} \max\left\{0, 1 - y^i\langle \boldsymbol{w}, \boldsymbol{x}^i\rangle\right\}}_{R(\boldsymbol{w})} \right)$$

Example: cutting plane model for SVM

- The convex problem to solver

$$\boldsymbol{w}^* \in \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \underbrace{\frac{1}{m}\sum_{i=1}^m \max\left\{0, 1 - y^i\langle \boldsymbol{w}, \boldsymbol{x}^i\rangle\right\}}_{R(\boldsymbol{w})} \right)$$

- The cutting plane model of $R(\boldsymbol{w})$ reads

$$R_t(\boldsymbol{w}) = \max_{i=0,\ldots,t-1} \Big[R(\boldsymbol{w}_i) + \langle \boldsymbol{g}_i, \boldsymbol{w} - \boldsymbol{w}_i \rangle \Big]$$

- The subgradient of $R(\boldsymbol{w})$ at $\boldsymbol{w}$

$$\boldsymbol{g}_i = -\frac{1}{m}\sum_{i=1}^m y^i \boldsymbol{x}^i \, [\![\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle \leq 1]\!]$$

Example: cutting plane model for SO-SVM with margin-rescaling loss

- The convex problem to solver

$$\boldsymbol{w}^* \in \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \left(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \underbrace{\frac{1}{m}\sum_{i=1}^{m} \max_{y\in\mathcal{Y}}\{\ell_i(y) + \langle \boldsymbol{w}, \boldsymbol{\phi}_i(y)\rangle\}}_{R(\boldsymbol{w})} \right)$$

Example: cutting plane model for SO-SVM with margin-rescaling loss

- The convex problem to solver

$$\boldsymbol{w}^* \in \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \bigg(\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \underbrace{\frac{1}{m}\sum_{i=1}^m \max_{y\in\mathcal{Y}}\{\ell_i(y) + \langle \boldsymbol{w}, \boldsymbol{\phi}_i(y)\rangle\}}_{R(\boldsymbol{w})} \bigg)$$

- The cutting plane model of $R(\boldsymbol{w})$ reads

$$R_t(\boldsymbol{w}) = \max_{i=0,\ldots,t-1} \Big[R(\boldsymbol{w}_i) + \langle \boldsymbol{g}_i, \boldsymbol{w} - \boldsymbol{w}_i \rangle \Big]$$

- The subgradient of $R(\boldsymbol{w})$ at $\boldsymbol{w}$

$$\boldsymbol{g}_i = \frac{1}{m}\sum_{i=1}^m \boldsymbol{\phi}_i(\hat{y}^i) \qquad \text{where} \qquad \hat{y}^i = \operatorname*{argmax}_{y\in\mathcal{Y}}\{\ell_i(y) + \langle \boldsymbol{w}_i, \boldsymbol{\phi}_i(y)\rangle\}$$

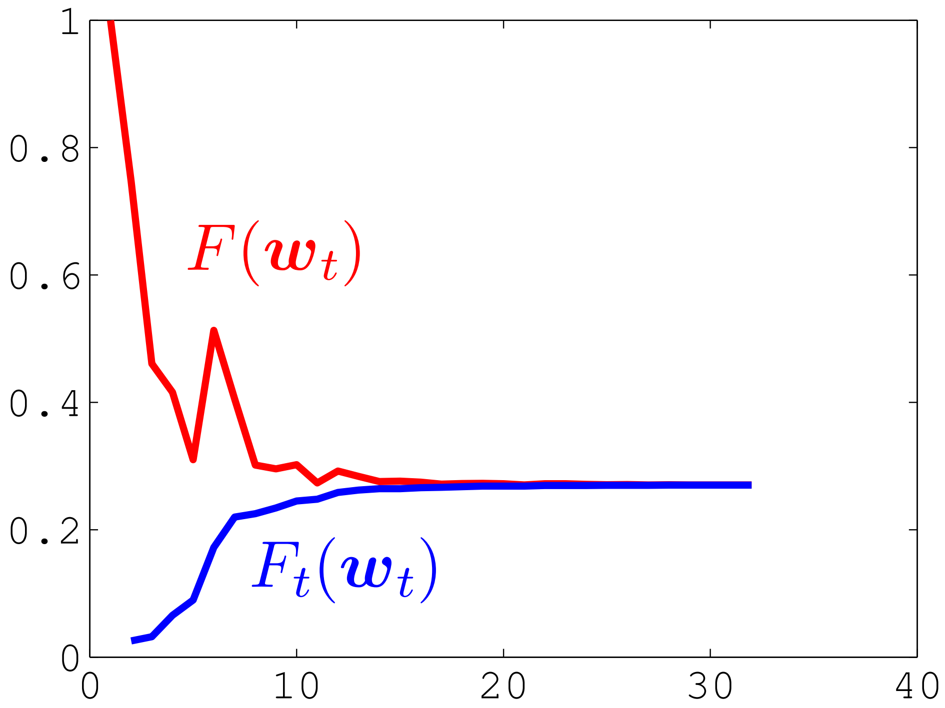
Bundle Method for Risk Minimization

1. Init: $t \leftarrow 0$, $\boldsymbol{w}_0 \in \mathbb{R}^n$
2. Compute $R(\boldsymbol{w}_t)$ and $\boldsymbol{g}_t \in \partial R(\boldsymbol{w}_t)$
3. $$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n} \left(\frac{\lambda}{2} \|\boldsymbol{w}\|^2 + R_t(\boldsymbol{w}) \right)$$

   where

   $$R_t(\boldsymbol{w}) = \max_{i=0,\ldots,t} \Big[R(\boldsymbol{w}_i) + \langle \boldsymbol{g}_i, \boldsymbol{w} - \boldsymbol{w}_i \rangle \Big]$$

4. if $\min_{i=1,\ldots,t} F(\boldsymbol{w}_i) - F_t(\boldsymbol{w}_{t+1}) \leq \varepsilon$ stop else $t \leftarrow t + 1$ go to 2.



How to solve the reduced problem

- Let us define a matrix $\mathbf{A} = [\boldsymbol{g}_0, \ldots, \boldsymbol{g}_t] \in \mathbb{R}^{n \times t}$ and a vector $b = [b_0, \ldots, b_{t-1}]$ with components $b_i = R(\boldsymbol{w}_i) - \langle \boldsymbol{g}_i, \boldsymbol{w}_i \rangle$.

- The **reduced problem** can be expressed as

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n, \xi \in \mathbb{R}} \left[\frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \xi \right] \quad \text{s.t.} \quad \xi \geq \langle \boldsymbol{w}, \boldsymbol{g}_i \rangle + b_i \, , i \in \{0, \ldots, t\}$$

How to solve the reduced problem

- Let us define a matrix $\mathbf{A} = [\boldsymbol{g}_0, \dots, \boldsymbol{g}_t] \in \mathbb{R}^{n \times t}$ and a vector $b = [b_0, \dots, b_{t-1}]$ with components $b_i = R(\boldsymbol{w}_i) - \langle \boldsymbol{g}_i, \boldsymbol{w}_i \rangle$.

- The **reduced problem** can be expressed as

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n, \xi \in \mathbb{R}} \left[\frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \xi \right] \quad \text{s.t.} \quad \xi \geq \langle \boldsymbol{w}, \boldsymbol{g}_i \rangle + b_i \,, i \in \{0, \dots, t\}$$

- The Lagrange **dual of the reduced problem** reads

$$\boldsymbol{\alpha}_{t+1} = \operatorname*{argmax}_{\boldsymbol{\alpha} \in \mathbb{R}^t} \left[\langle \boldsymbol{\alpha}, \boldsymbol{b} \rangle - \frac{1}{2\lambda} \langle \boldsymbol{\alpha}, \mathbf{A}^T \mathbf{A} \boldsymbol{\alpha} \rangle \right] \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_1 = 1, \boldsymbol{\alpha} \geq \mathbf{0}$$

How to solve the reduced problem

- Let us define a matrix $\mathbf{A} = [\boldsymbol{g}_0, \ldots, \boldsymbol{g}_t] \in \mathbb{R}^{n \times t}$ and a vector $b = [b_0, \ldots, b_{t-1}]$ with components $b_i = R(\boldsymbol{w}_i) - \langle \boldsymbol{g}_i, \boldsymbol{w}_i \rangle$.

- The **reduced problem** can be expressed as

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n, \xi \in \mathbb{R}} \left[\frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \xi \right] \quad \text{s.t.} \quad \xi \geq \langle \boldsymbol{w}, \boldsymbol{g}_i \rangle + b_i \, , i \in \{0, \ldots, t\}$$

- The Lagrange **dual of the reduced problem** reads

$$\boldsymbol{\alpha}_{t+1} = \operatorname*{argmax}_{\boldsymbol{\alpha} \in \mathbb{R}^t} \left[\langle \boldsymbol{\alpha}, \boldsymbol{b} \rangle - \frac{1}{2\lambda} \langle \boldsymbol{\alpha}, \mathbf{A}^T \mathbf{A} \boldsymbol{\alpha} \rangle \right] \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_1 = 1, \boldsymbol{\alpha} \geq \mathbf{0}$$

- The primal solution is $\boldsymbol{w}_{t+1} = -\lambda^{-1} \mathbf{A} \boldsymbol{\alpha}_{t+1}$

Bundle Method for Risk Minimization

(+) Provides a certificate of optimility:

$$F(\boldsymbol{w}_t) \leq F(\boldsymbol{w}^*) + \varepsilon$$

Bundle Method for Risk Minimization

(+) Provides a certificate of optimility:

$$F(\boldsymbol{w}_t) \leq F(\boldsymbol{w}^*) + \varepsilon$$

(+) Converges for arbitrary $\varepsilon > 0$ in

$$\log_2 \frac{\lambda F(\mathbf{0})}{G^2} + \frac{4G^2}{\lambda \varepsilon} - 1$$

iteration at most where $G \geq \|\boldsymbol{g}\|_2$, $\forall \boldsymbol{g} \in \partial R(\boldsymbol{w}), \boldsymbol{w} \in \mathbb{R}^n$.

Bundle Method for Risk Minimization

(+) Provides a certificate of optimility:

$$F(\boldsymbol{w}_t) \leq F(\boldsymbol{w}^*) + \varepsilon$$

(+) Converges for arbitrary $\varepsilon > 0$ in

$$\log_2 \frac{\lambda F(\mathbf{0})}{G^2} + \frac{4G^2}{\lambda \varepsilon} - 1$$

iteration at most where $G \geq \|\boldsymbol{g}\|_2$, $\forall \boldsymbol{g} \in \partial R(\boldsymbol{w}), \boldsymbol{w} \in \mathbb{R}^n$.

(+) Requires only the first order oracle computing $R(\boldsymbol{w})$ and $\boldsymbol{g} \in \partial R(\boldsymbol{w})$.

Bundle Method for Risk Minimization

(+) Provides a certificate of optimility:

$$F(\boldsymbol{w}_t) \leq F(\boldsymbol{w}^*) + \varepsilon$$

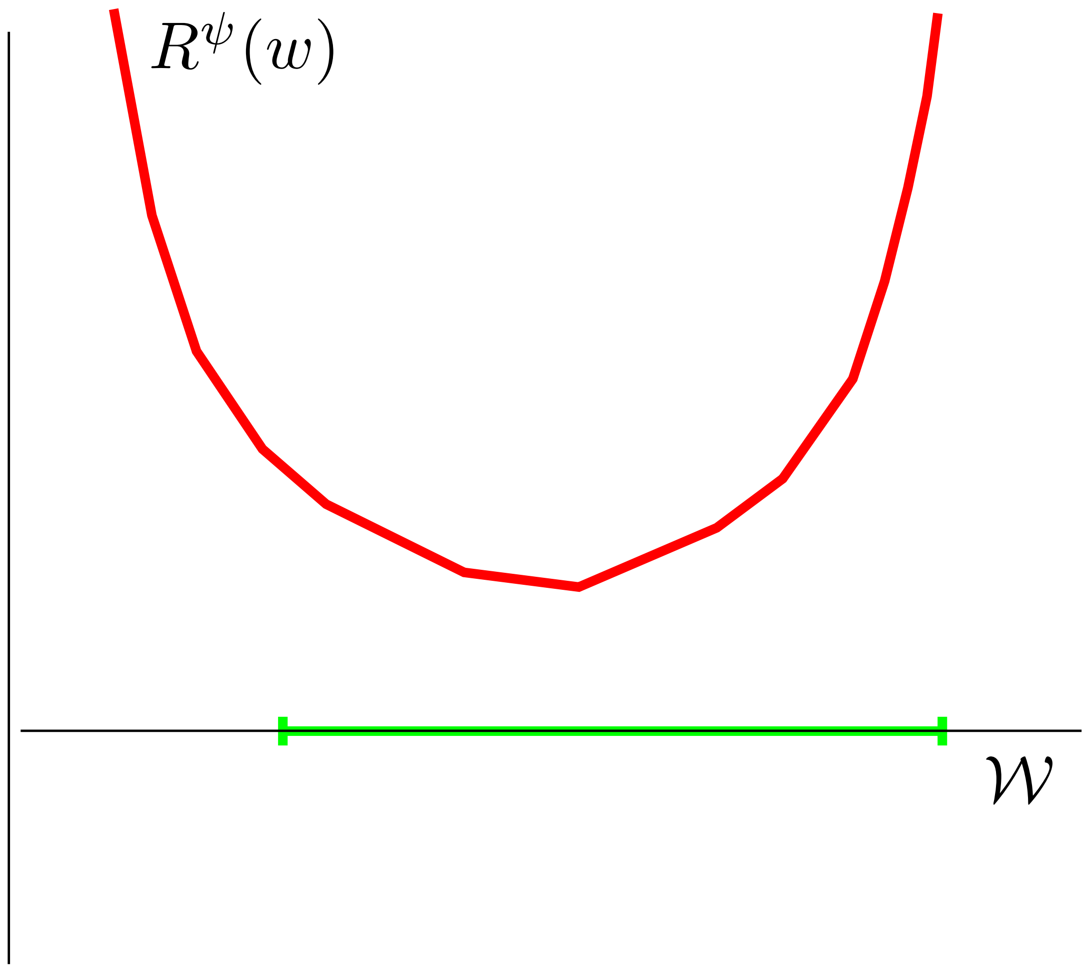
(+) Converges for arbitrary $\varepsilon > 0$ in

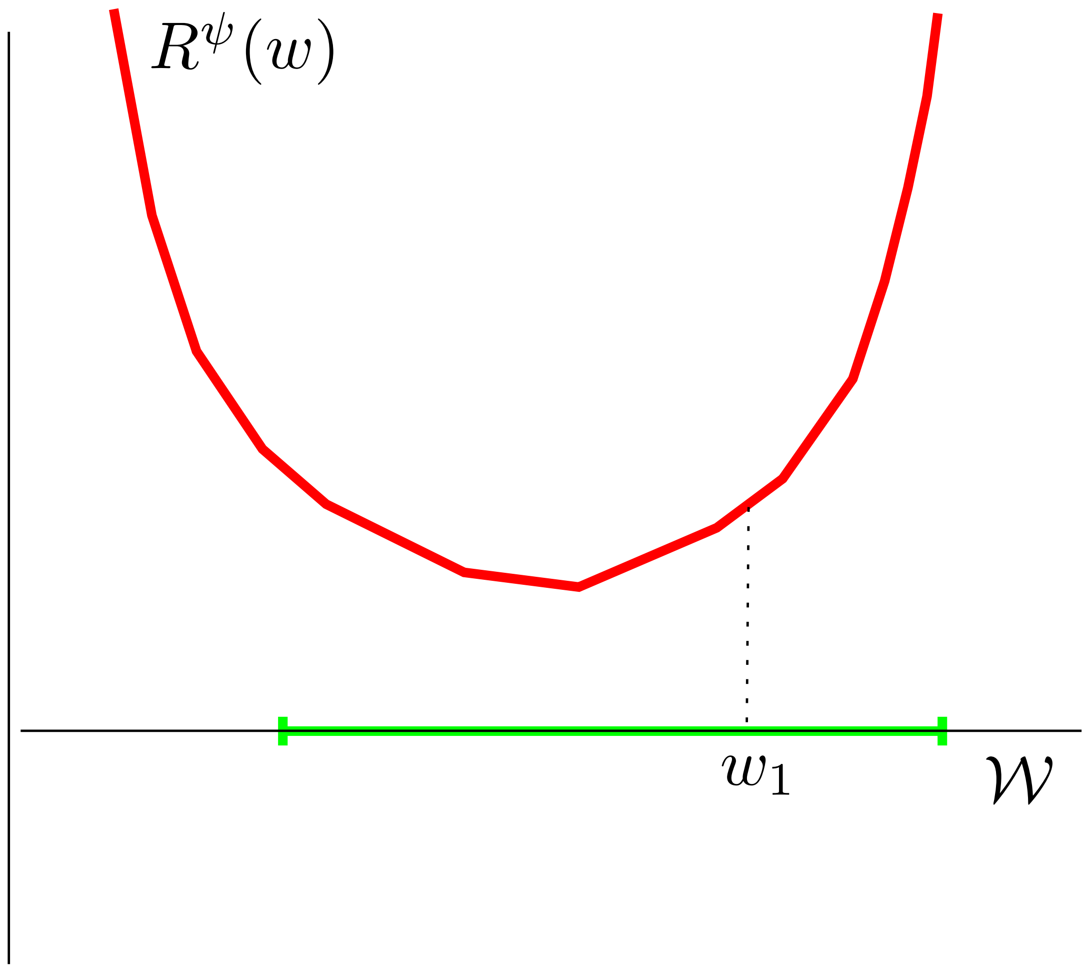
$$\log_2 \frac{\lambda F(\mathbf{0})}{G^2} + \frac{4G^2}{\lambda\,\varepsilon} - 1$$

iteration at most where $G \geq \|\boldsymbol{g}\|_2$, $\forall \boldsymbol{g} \in \partial R(\boldsymbol{w})$,$\boldsymbol{w} \in \mathbb{R}^n$.

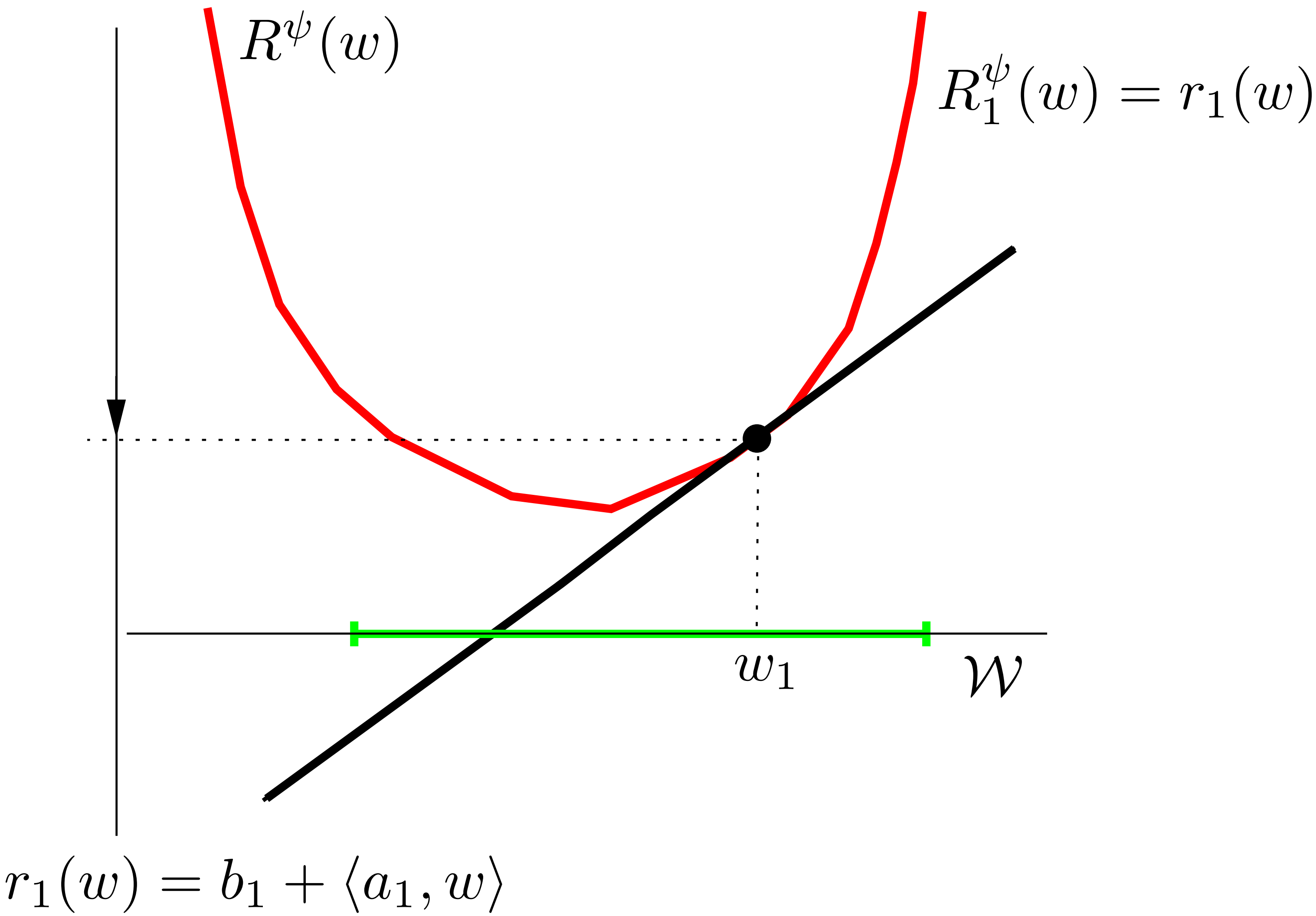
(+) Requires only the first order oracle computing $R(\boldsymbol{w})$ and $\boldsymbol{g} \in \partial R(\boldsymbol{w})$.

(-) Slow convergence for $\lambda \to 0$.

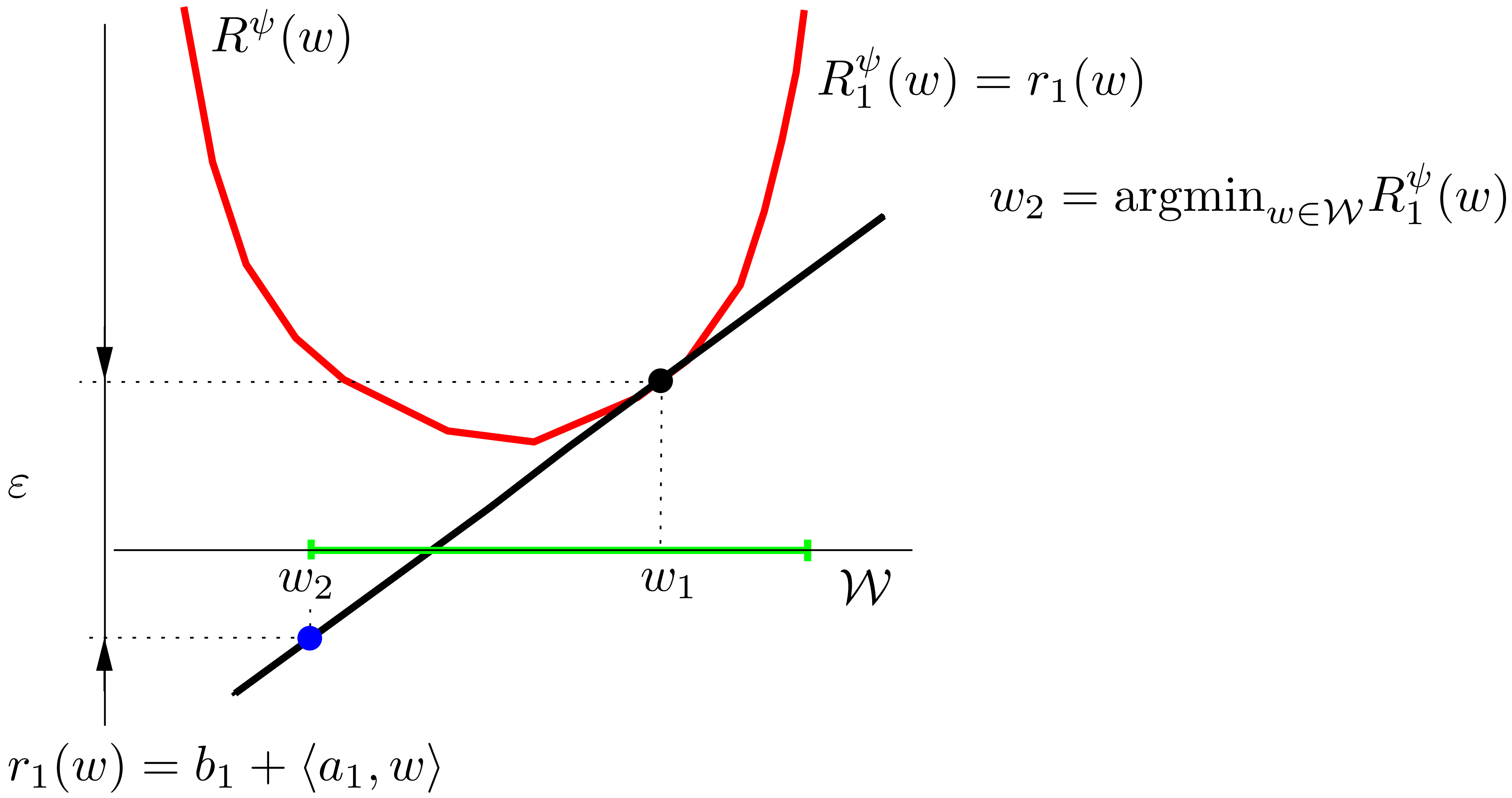

$R^{\psi}(w)$
$\mathcal{W}$



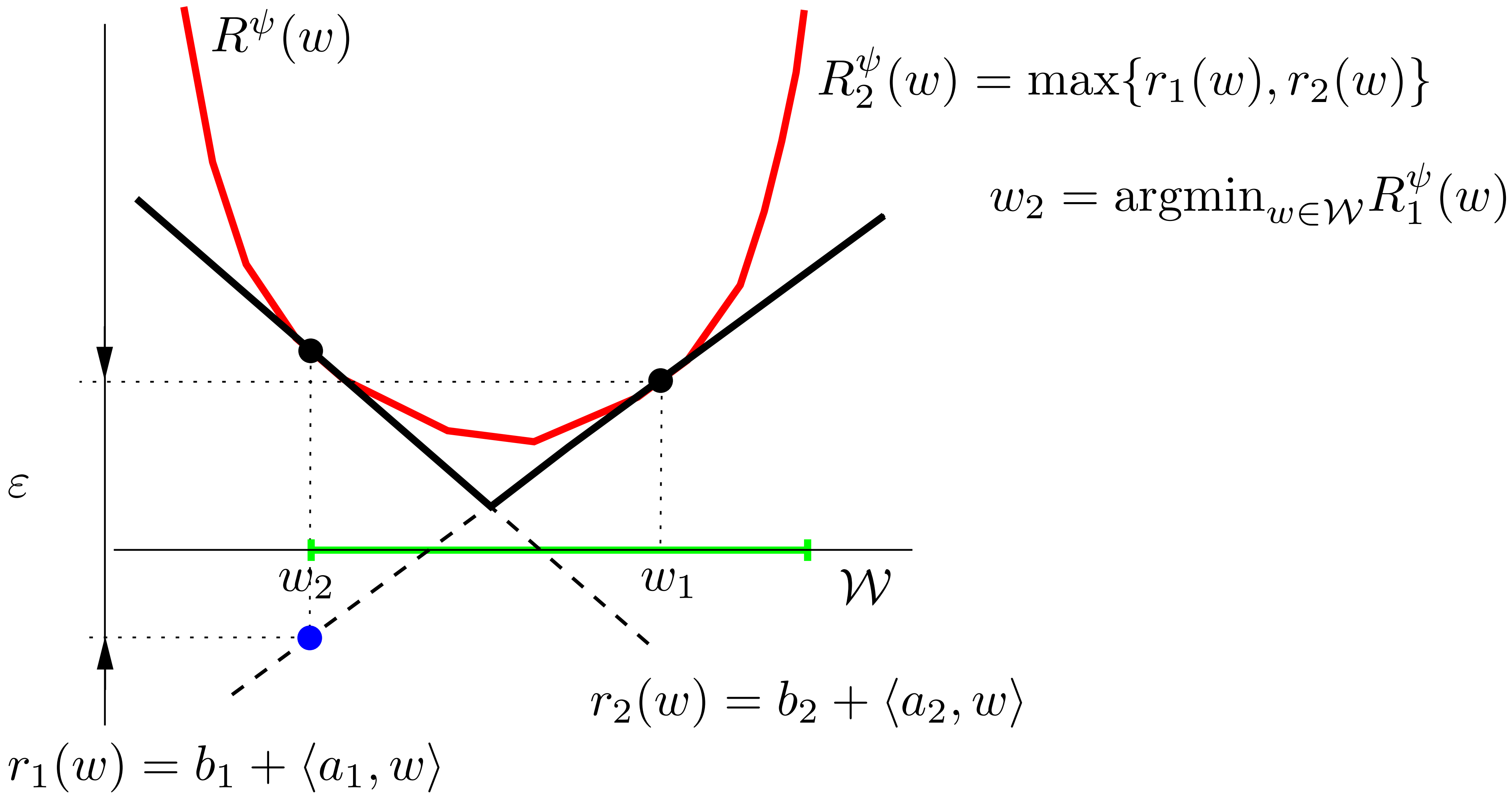
$R^{\psi}(w)$
w_1
$\mathcal{W}$



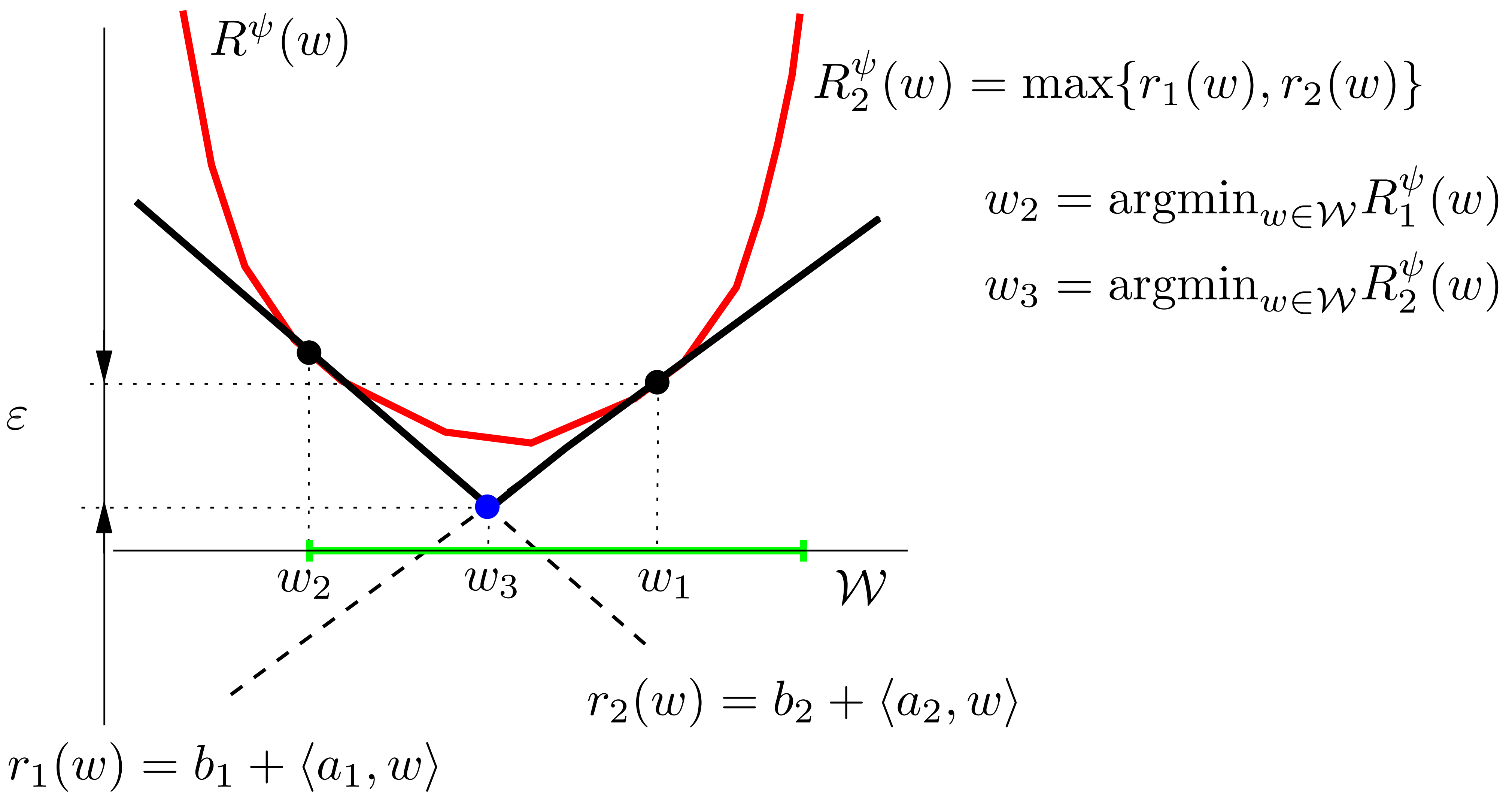
$R^{\psi}(w)$
$R_1^{\psi}(w) = r_1(w)$
w_1
$\mathcal{W}$
$r_1(w) = b_1 + \langle a_1, w \rangle$

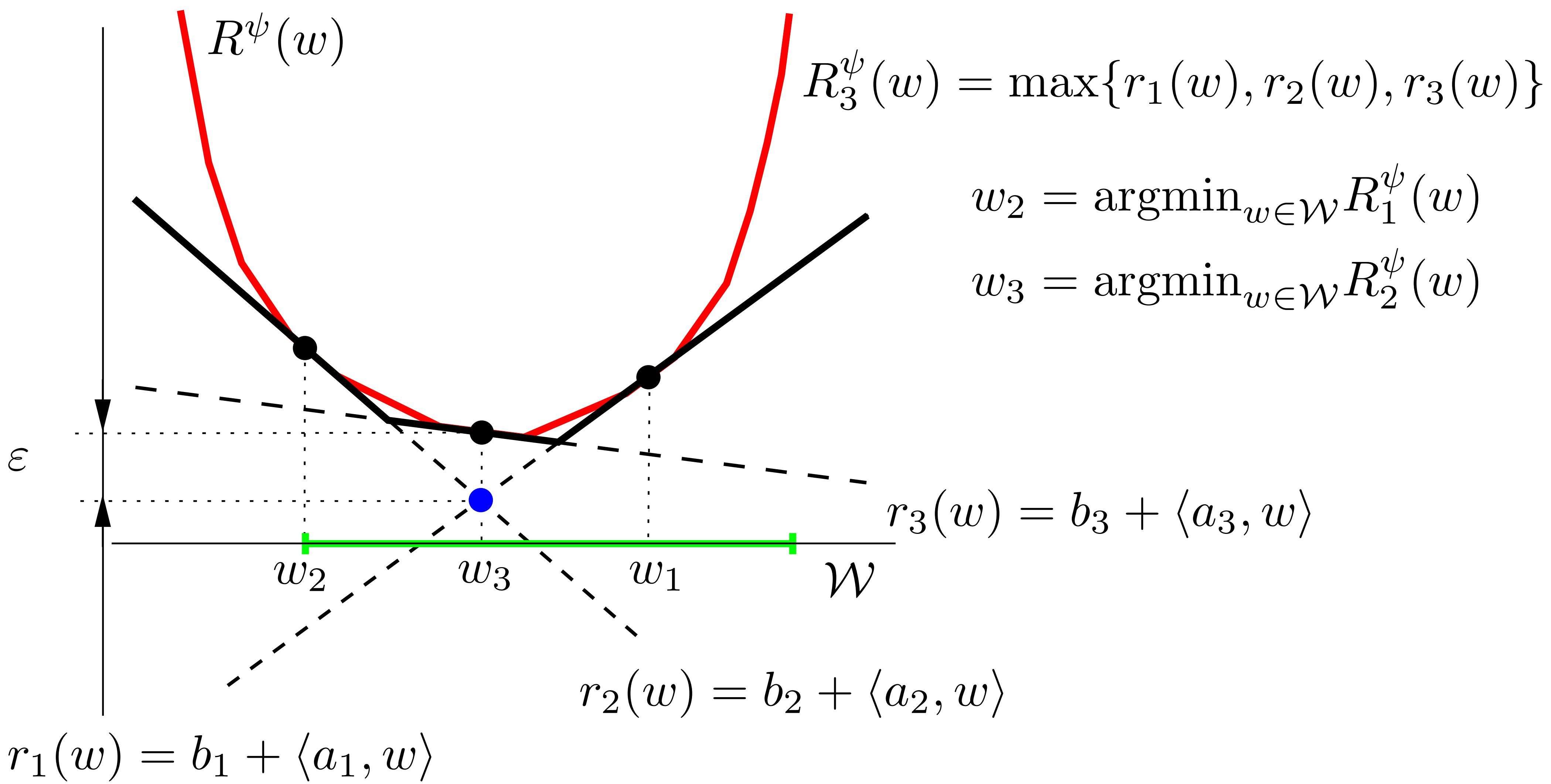


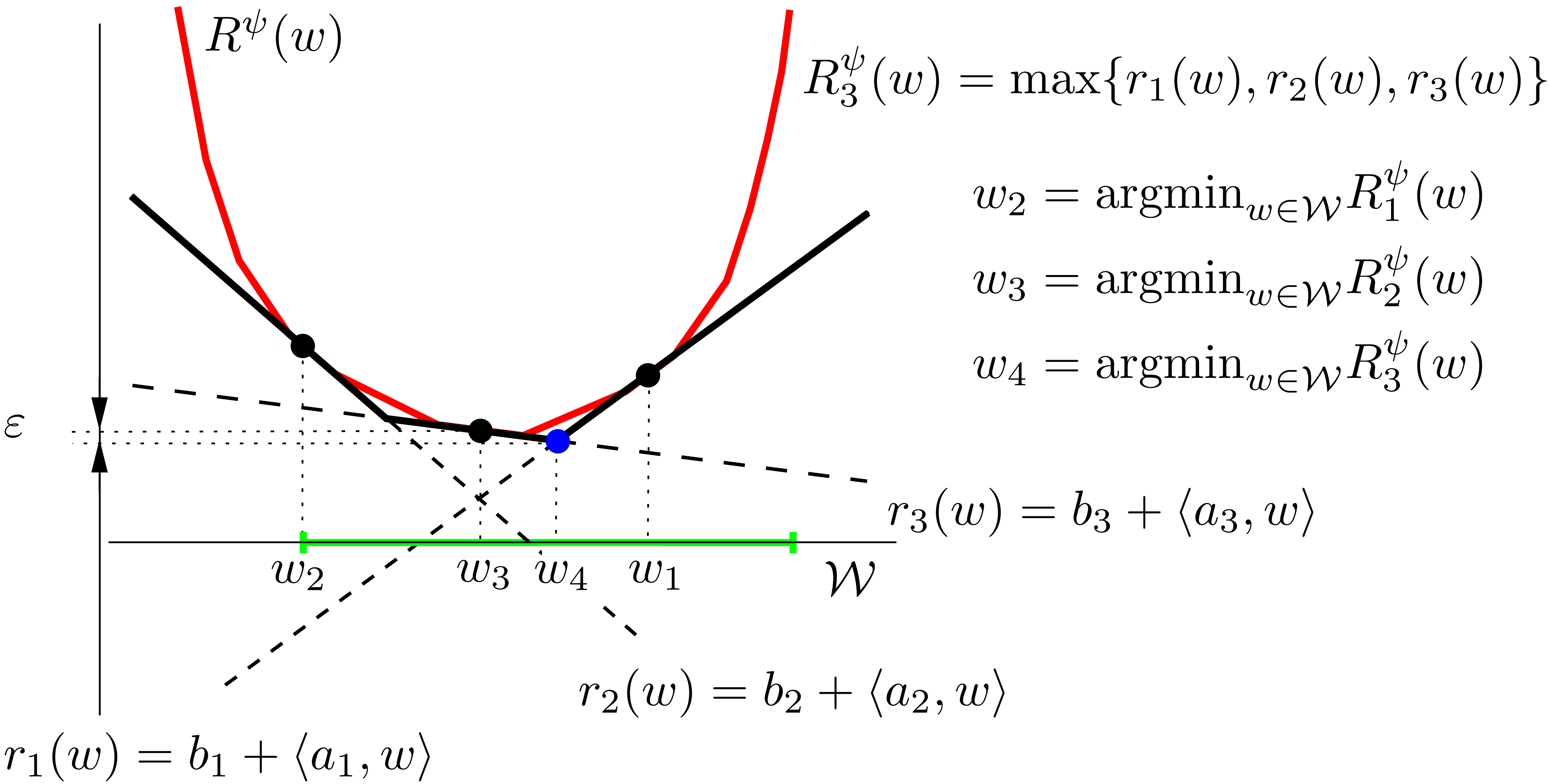
$R^{\psi}(w)$
$R_1^{\psi}(w) = r_1(w)$
$w_2 = \mathrm{argmin}_{w \in \mathcal{W}} R_1^{\psi}(w)$
ε
w_2
w_1
$\mathcal{W}$
$r_1(w) = b_1 + \langle a_1, w \rangle$

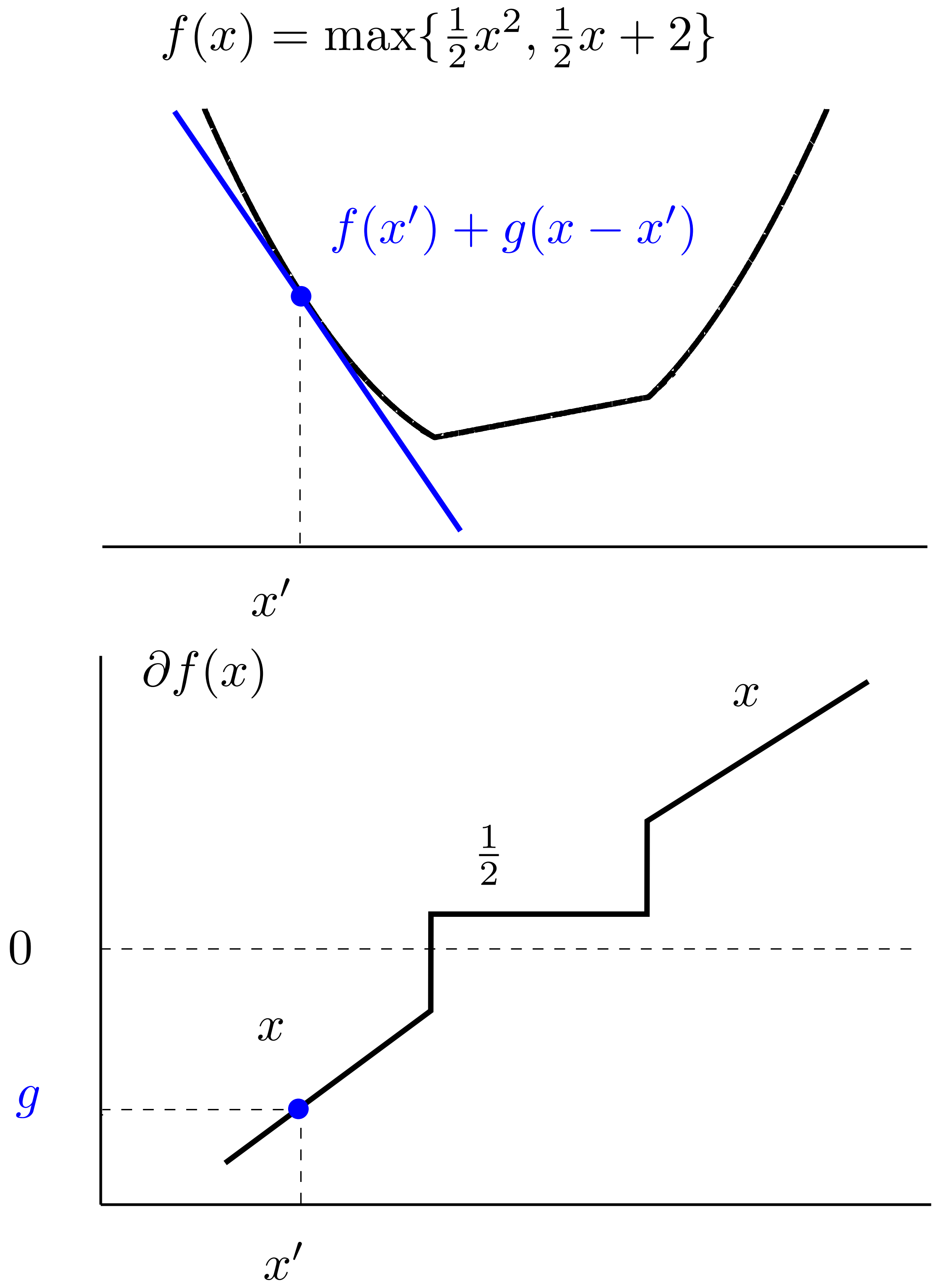


$R^{\psi}(w)$
$R_2^{\psi}(w) = \max\{r_1(w), r_2(w)\}$
$w_2 = \mathrm{argmin}_{w \in \mathcal{W}} R_1^{\psi}(w)$
ε
w_2
w_1
$\mathcal{W}$
$r_2(w) = b_2 + \langle a_2, w \rangle$
$r_1(w) = b_1 + \langle a_1, w \rangle$


$R^{\psi}(w)$
$R^{\psi}_2(w) = \max\{r_1(w), r_2(w)\}$
$w_2 = \text{argmin}_{w \in \mathcal{W}} R^{\psi}_1(w)$
$w_3 = \text{argmin}_{w \in \mathcal{W}} R^{\psi}_2(w)$
ε
w_2
w_3
w_1
$\mathcal{W}$
$r_2(w) = b_2 + \langle a_2, w \rangle$
$r_1(w) = b_1 + \langle a_1, w \rangle$

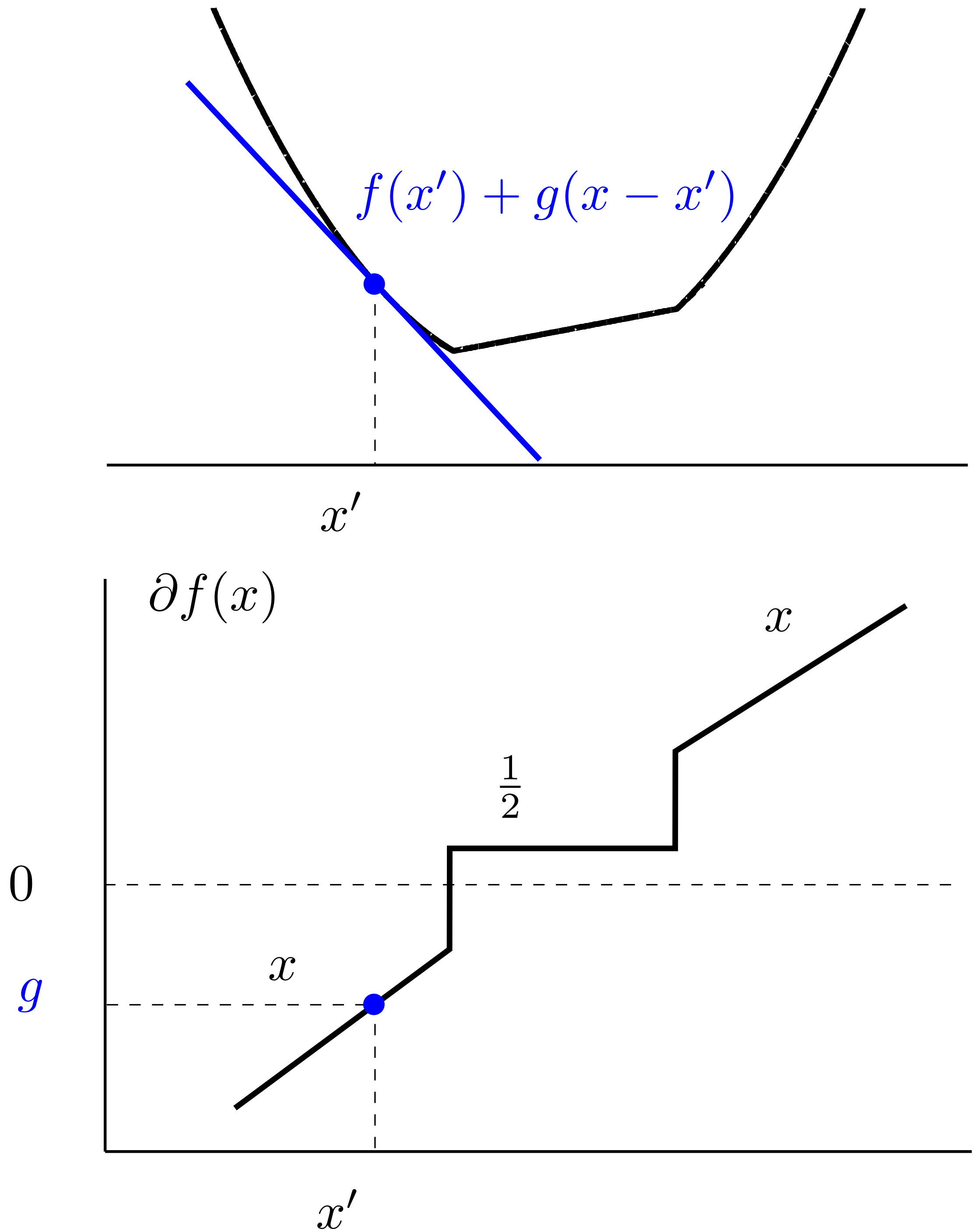

$R^\psi(w)$
$R_3^\psi(w) = \max\{r_1(w), r_2(w), r_3(w)\}$
$w_2 = \mathrm{argmin}_{w\in\mathcal{W}} R_1^\psi(w)$
$w_3 = \mathrm{argmin}_{w\in\mathcal{W}} R_2^\psi(w)$
ε
$r_3(w) = b_3 + \langle a_3, w\rangle$
w_2
w_3
w_1
$\mathcal{W}$
$r_2(w) = b_2 + \langle a_2, w\rangle$
$r_1(w) = b_1 + \langle a_1, w\rangle$


$R^{\psi}(w)$
$R^{\psi}_3(w) = \max\{r_1(w), r_2(w), r_3(w)\}$
$w_2 = \mathrm{argmin}_{w \in \mathcal{W}} R^{\psi}_1(w)$
$w_3 = \mathrm{argmin}_{w \in \mathcal{W}} R^{\psi}_2(w)$
$w_4 = \mathrm{argmin}_{w \in \mathcal{W}} R^{\psi}_3(w)$
ε
$r_3(w) = b_3 + \langle a_3, w \rangle$
w_2
w_3
w_4
w_1
$\mathcal{W}$
$r_2(w) = b_2 + \langle a_2, w \rangle$
$r_1(w) = b_1 + \langle a_1, w \rangle$

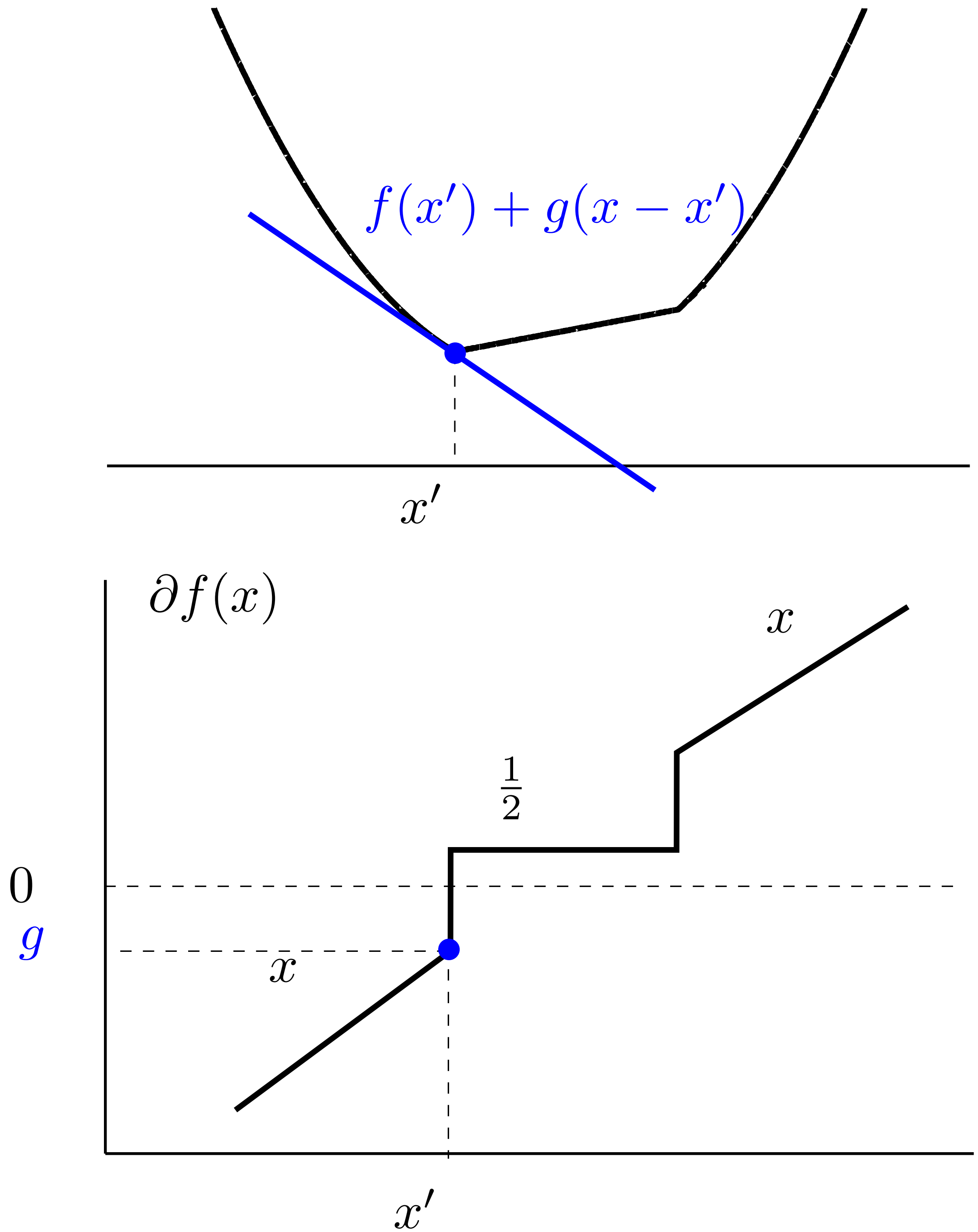


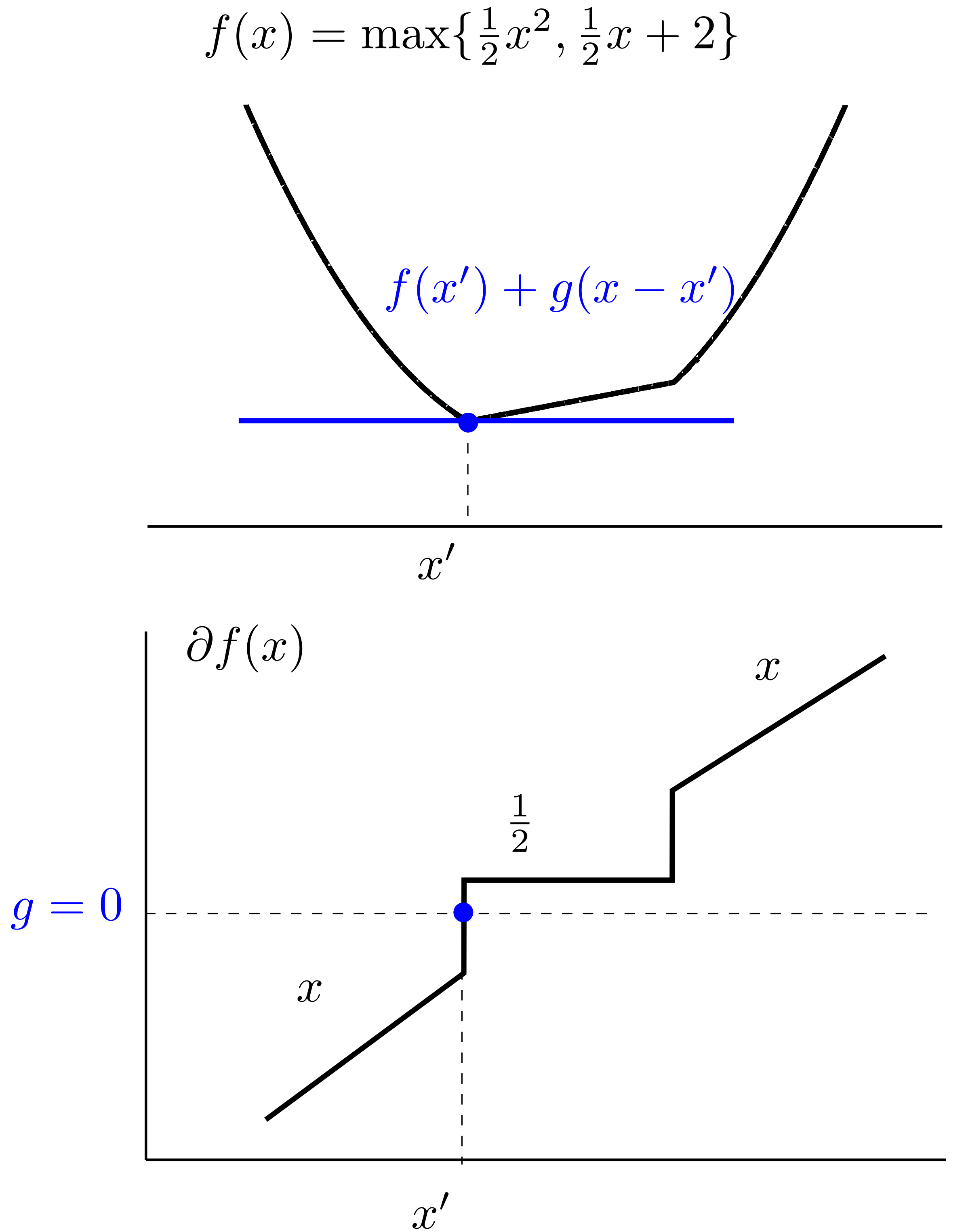
$f(x) = \max\{\frac{1}{2}x^2, \frac{1}{2}x + 2\}$
$f(x') + g(x - x')$
x'
$\partial f(x)$
x
$\frac{1}{2}$
0
x
g
x'

$$f(x) = \max\{\tfrac{1}{2}x^2, \tfrac{1}{2}x + 2\}$$

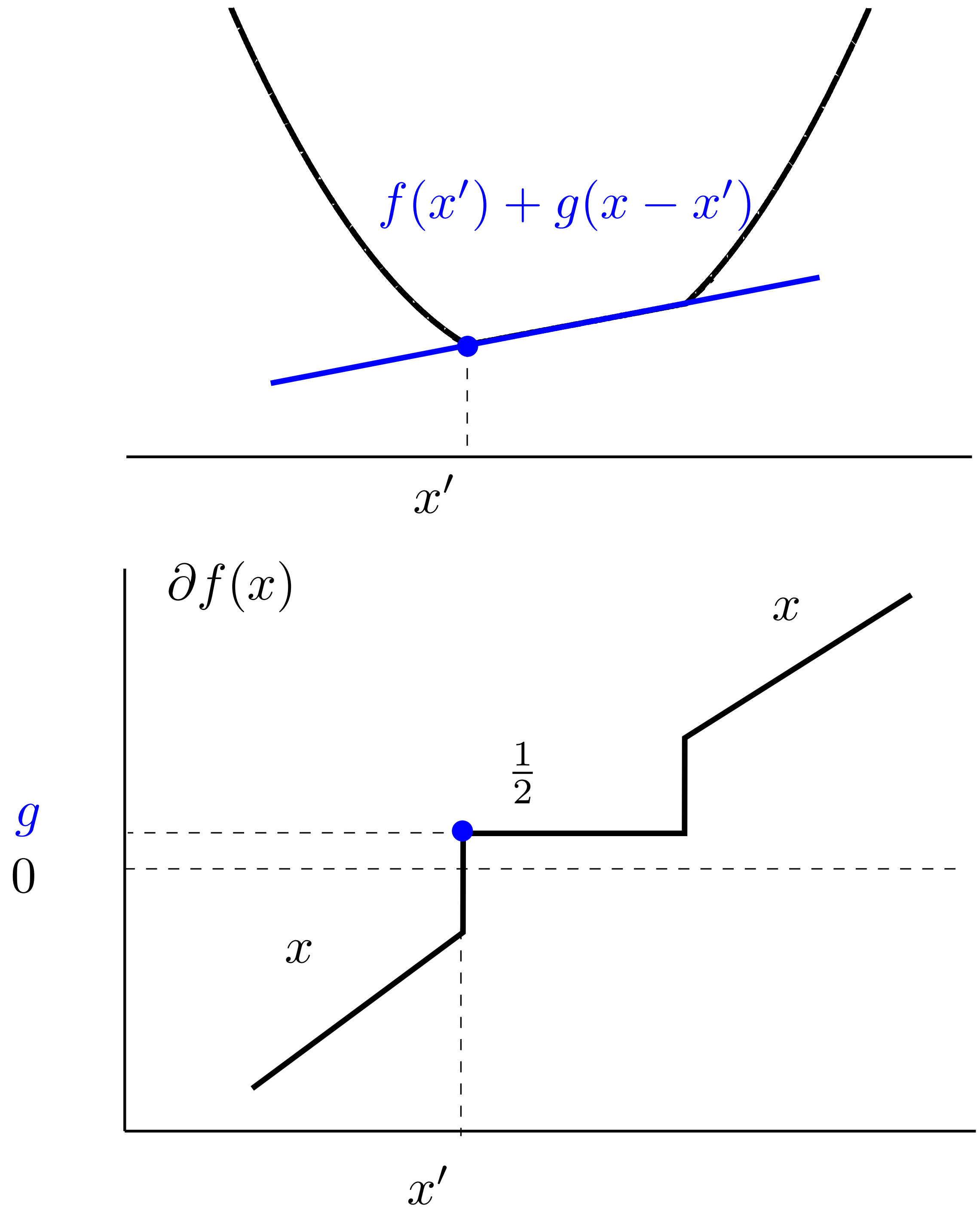


$$f(x) = \max\{\tfrac{1}{2}x^2, \tfrac{1}{2}x + 2\}$$

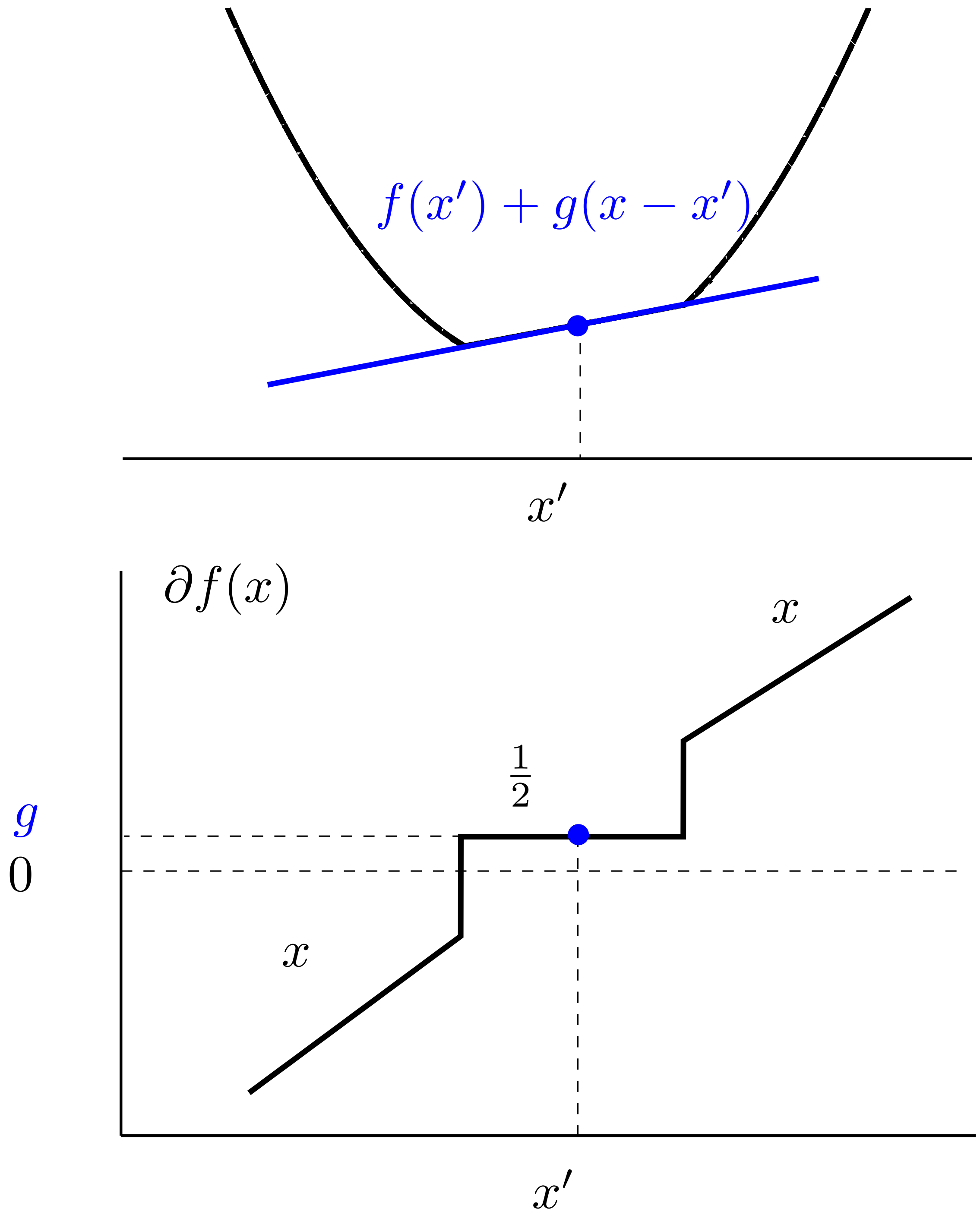



$f(x) = \max\{\frac{1}{2}x^2, \frac{1}{2}x + 2\}$
$f(x') + g(x - x')$
x'
$\partial f(x)$
x
$\frac{1}{2}$
$g = 0$
x
x'

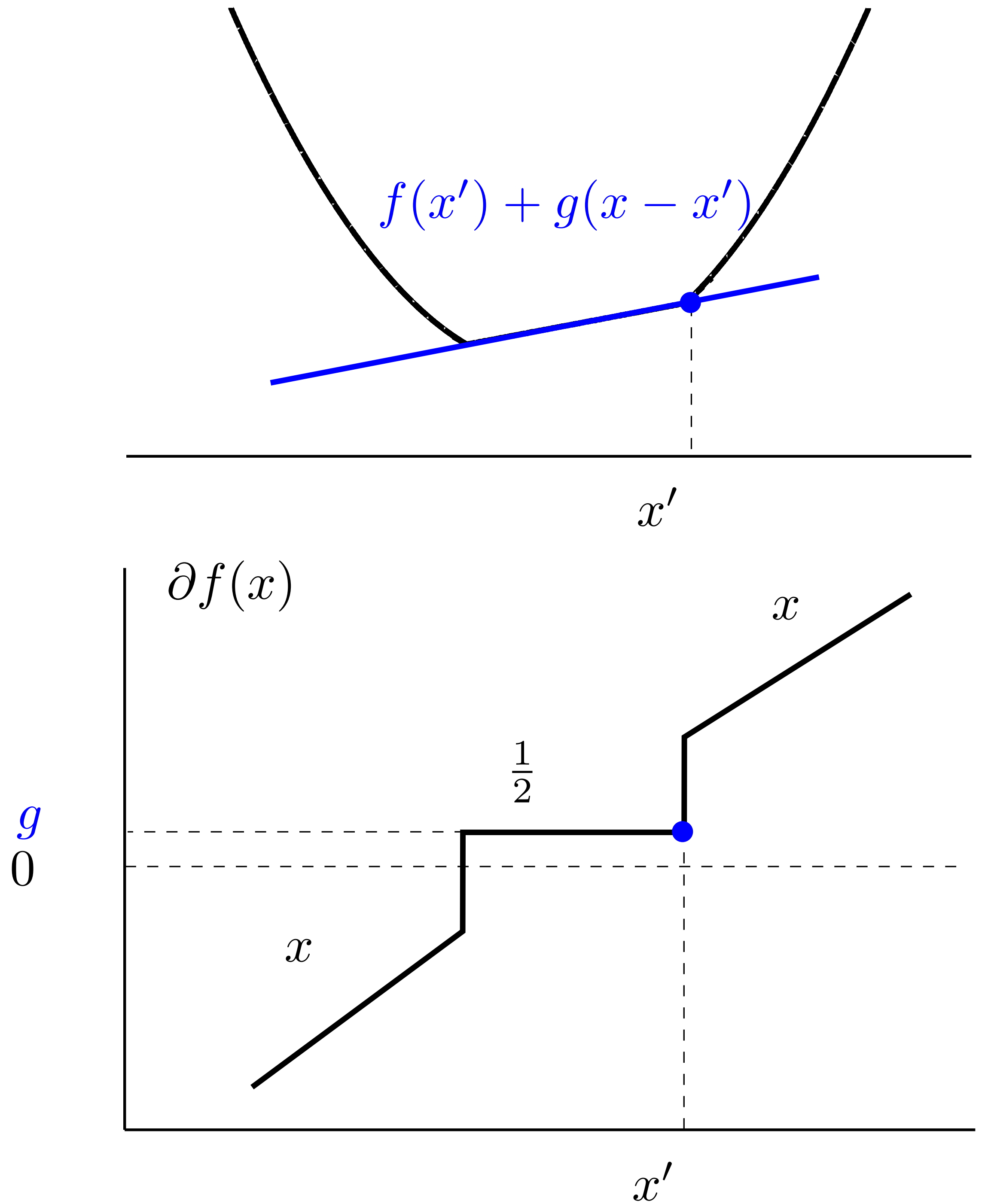
$$f(x) = \max\{\tfrac{1}{2}x^2, \tfrac{1}{2}x + 2\}$$



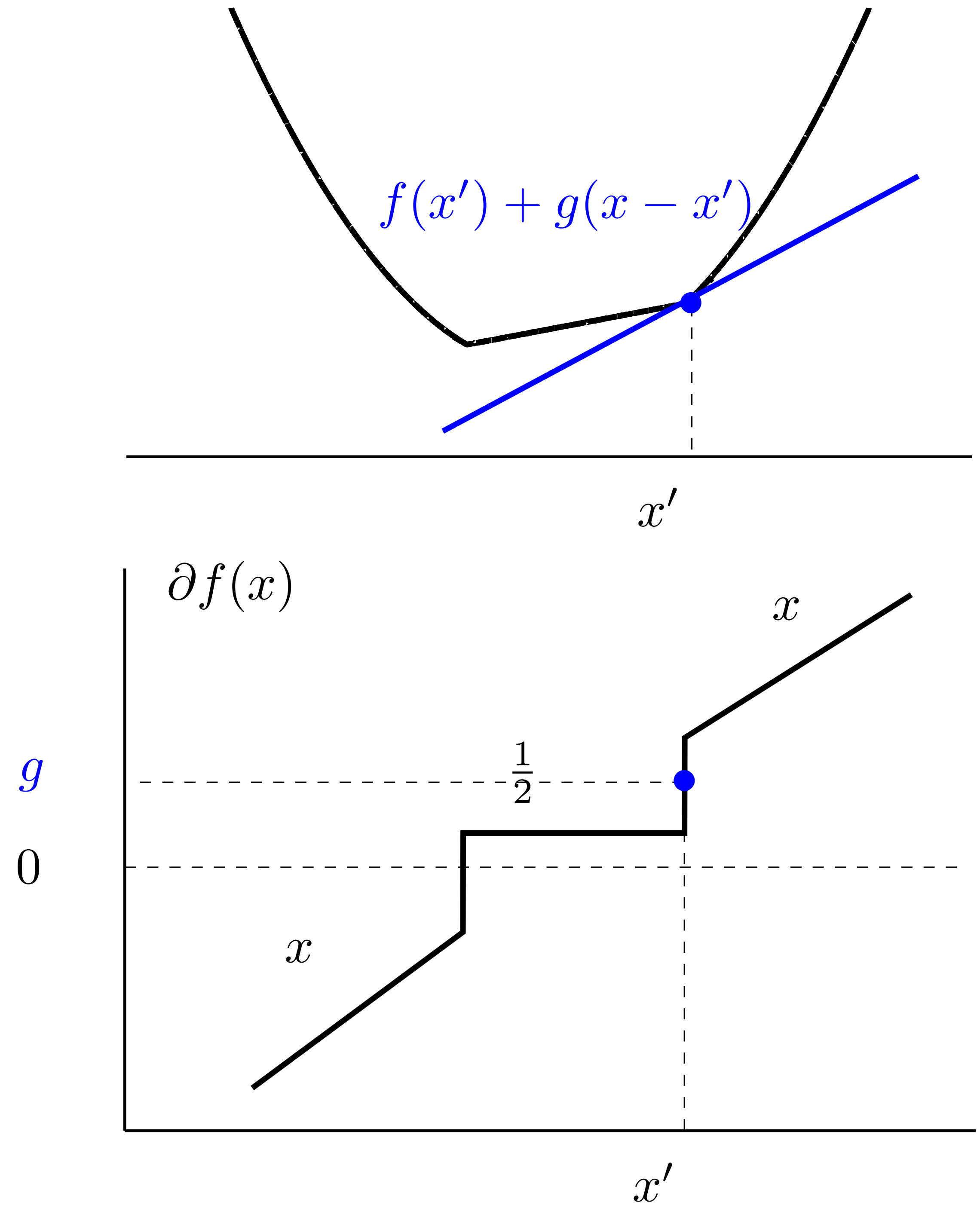
$$f(x) = \max\{\tfrac{1}{2}x^2, \tfrac{1}{2}x + 2\}$$

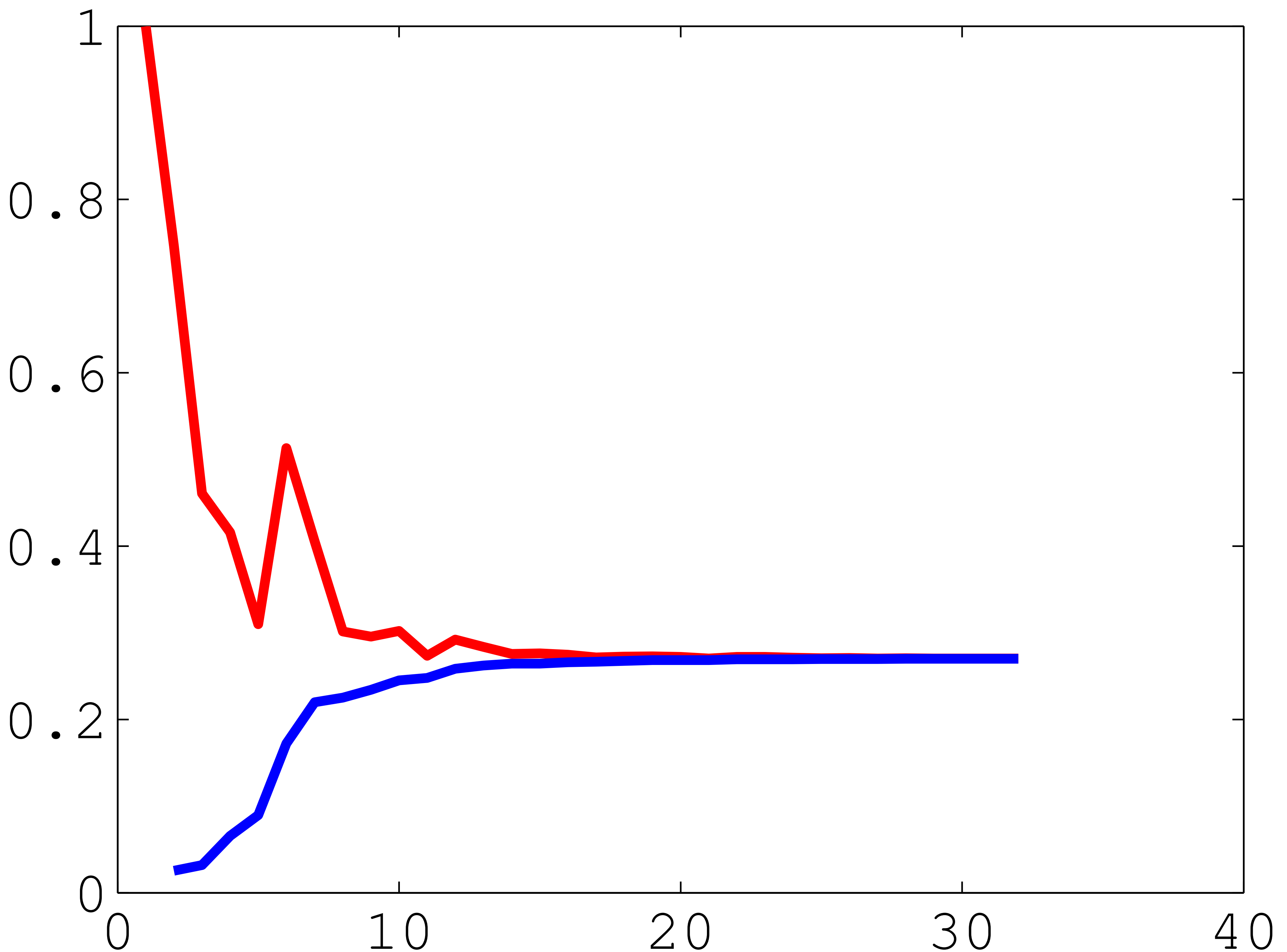


$$f(x) = \max\{\tfrac{1}{2}x^2, \tfrac{1}{2}x + 2\}$$



$$f(x) = \max\{\tfrac{1}{2}x^2, \tfrac{1}{2}x + 2\}$$




1
0.8
0.6
0.4
0.2
0
0
10
20
30
40