Poster #212

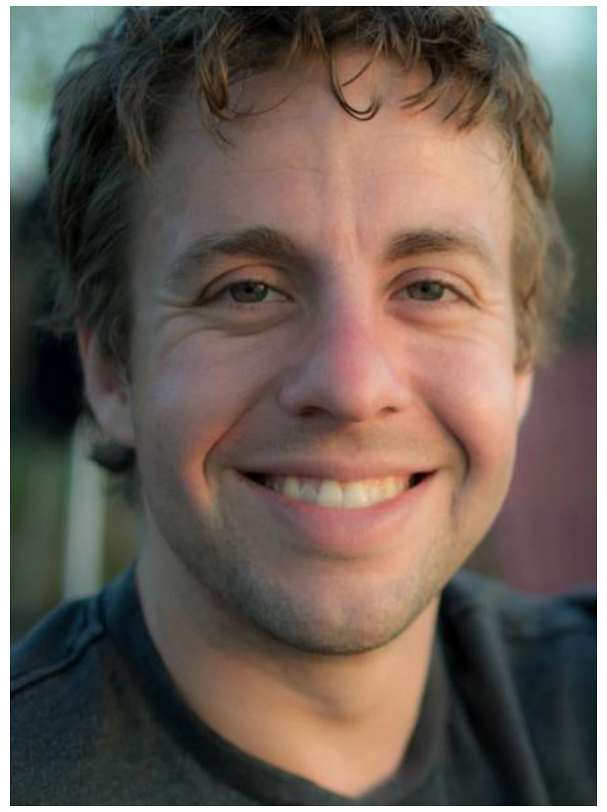
Variance Reduction for Matrix Games

Yair Carmon

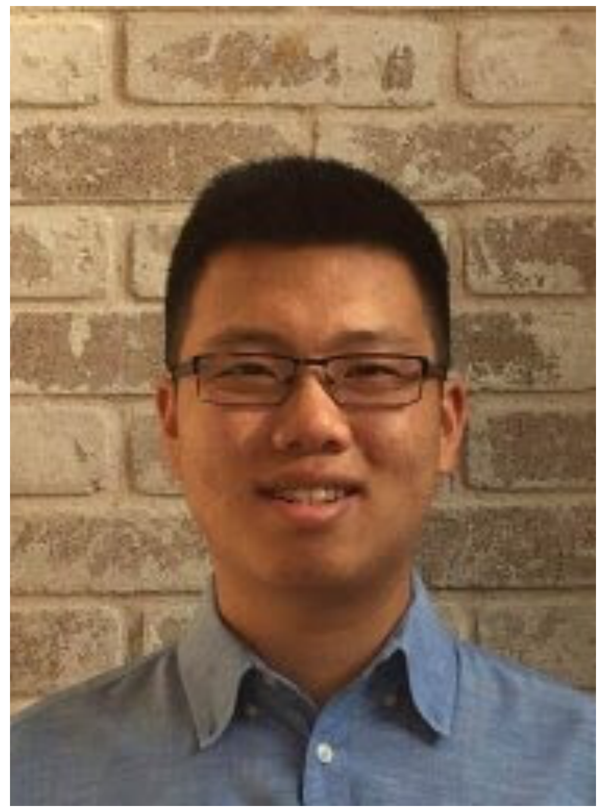
(presenting)

Yujia Jin

Aaron Sidford



Kevin Tian



Stanford University

Zero-sum games

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$$

Super useful!

- Constraints: y checks feasibility (e.g. GAN's)
- Robustness: y represents uncertainty (e.g. adversarial training)

Zero-sum games

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$$

Super useful!

- Constraints: y checks feasibility (e.g. GAN's)
- Robustness: y represents uncertainty (e.g. adversarial training)

Ideal (approximate) solution ϵ-Nash equilibrium

x player is happy

$$f(x, y) \leq \min_{x' \in \mathcal{X}} f(x', y) + \epsilon$$

y player is happy

$$f(x, y) \geq \max_{y' \in \mathcal{Y}} f(x, y') - \epsilon$$

Zero-sum games

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$$

Super useful!

- Constraints: y checks feasibility (e.g. GAN's)
- Robustness: y represents uncertainty (e.g. adversarial training)

Ideal (approximate) solution ϵ-Nash equilibrium

$$\overbrace{f(x, y) \le \min_{x' \in \mathcal{X}} f(x', y) + \epsilon}^{x \text{ player is happy}} \qquad \overbrace{f(x, y) \ge \max_{y' \in \mathcal{Y}} f(x, y') - \epsilon}^{y \text{ player is happy}}$$

We assume f is convex-concave $\Longrightarrow$ Nash equilibrium exists

Our contributions

Our contributions

1. Variance reduction framework for general (convex-concave) $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$

Our contributions

1. Variance reduction framework
   for general (convex-concave) $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$

Centered *gradient estimator* $\Longrightarrow$ Fast algorithm

Our contributions

1. Variance reduction framework
   for general (convex-concave) $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$

GEOMETRY MATTERS

Centered
gradient estimator ⟹ Fast algorithm

Poster #212

Our contributions

Geometry matters

1. Variance reduction framework
   for general (convex-concave) $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$

Centered *gradient estimator* $\Rightarrow$ Fast algorithm

2. Concrete centered gradient estimators
   for $f(x) = y^\top A x$

"Sampling from the difference"

Our contributions

1. Variance reduction framework for general (convex-concave) $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$

Centered gradient estimator $\Rightarrow$ Fast algorithm

2. Concrete centered gradient estimators for $f(x) = y^\top A x$

"Sampling from the difference"

GEOMETRY MATTERS

New runtimes for $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top A x$

Bilinear games

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top A x, \quad A \in \mathbb{R}^{m \times n}$$

- Simplest case
- Local model for smooth zero-sum game
- Important by themselves

Bilinear games

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top A x, \quad A \in \mathbb{R}^{m \times n}$$

- Simplest case
- Local model for smooth zero-sum game
- Important by themselves

GEOMETRY MATTERS

Bilinear games

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top A x, \quad A \in \mathbb{R}^{m \times n}$$

- Simplest case
- Local model for smooth zero-sum game
- Important by themselves

GEOMETRY MATTERS

$\mathcal{X} = \mathcal{Y} =$ simplex
Matrix games / LP

Bilinear games

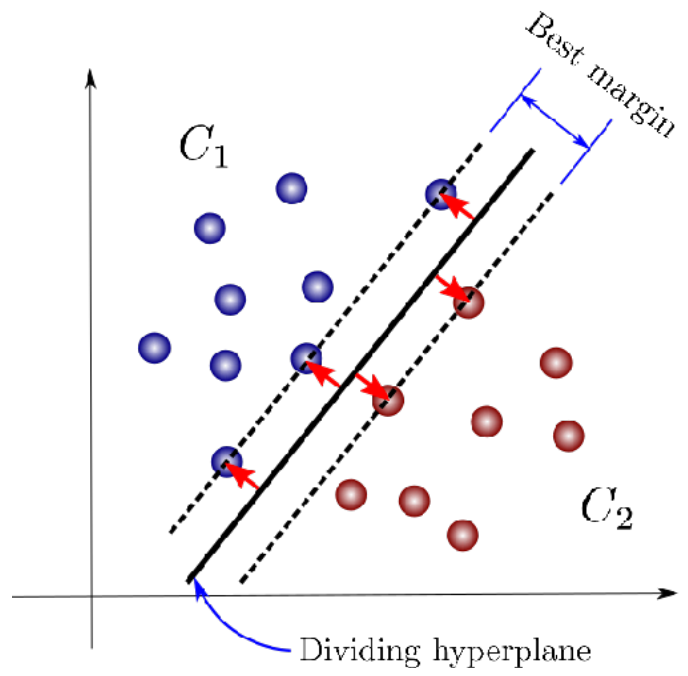
$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top A x, \quad A \in \mathbb{R}^{m \times n}$$

- Simplest case
- Local model for smooth zero-sum game
- Important by themselves

GEOMETRY MATTERS

$\mathcal{X} = \mathcal{Y} =$ simplex
Matrix games / LP

$\mathcal{X} =$ Euclidean, $\mathcal{Y} =$ simplex
Hard margin SVM



Bilinear games

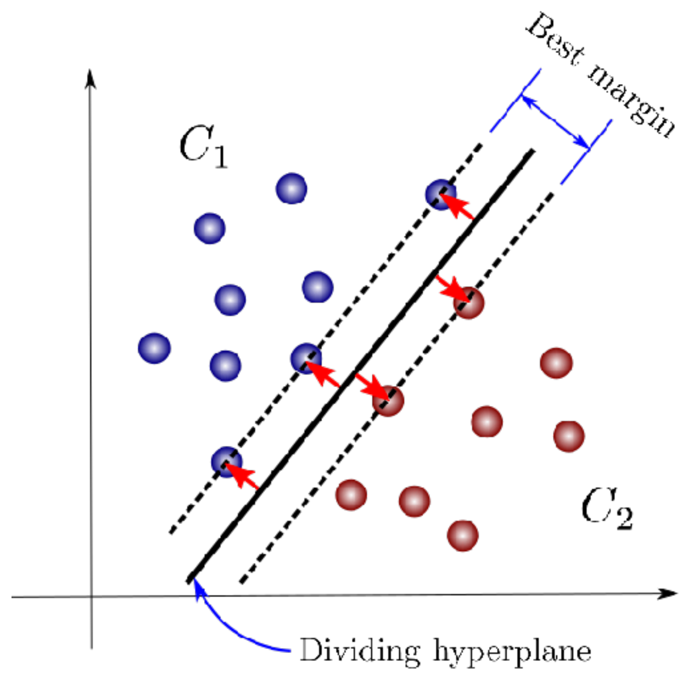
$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top A x, \quad A \in \mathbb{R}^{m \times n}$$

- Simplest case
- Local model for smooth zero-sum game
- Important by themselves

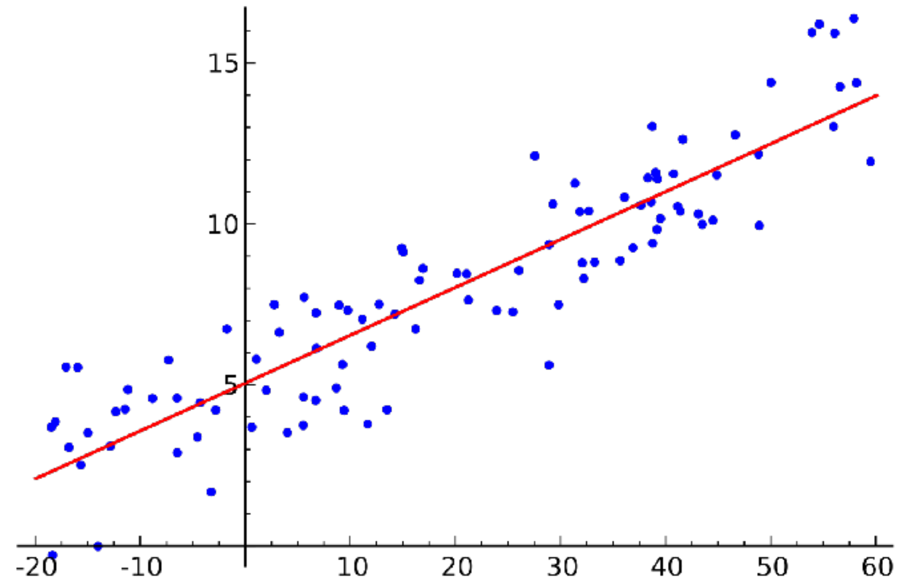
GEOMETRY MATTERS

$\mathcal{X} = \mathcal{Y} =$ simplex
Matrix games / LP

$\mathcal{X} =$ Euclidean, $\mathcal{Y} =$ simplex
Hard margin SVM



$\mathcal{X} = \mathcal{Y} =$ Euclidean
Linear regression



Poster #212

Bilinear games

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top A x, \quad A \in \mathbb{R}^{m \times n}$$

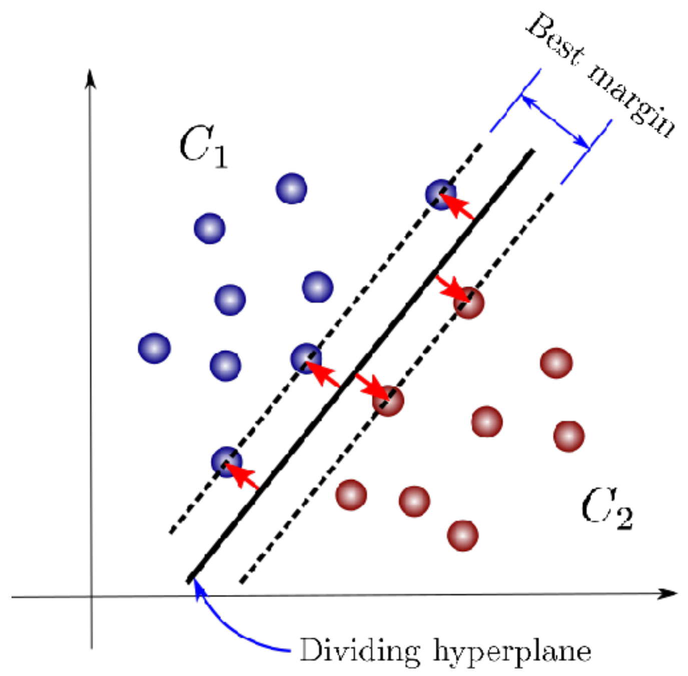
- Simplest case
- Local model for smooth zero-sum game
- Important by themselves

GEOMETRY MATTERS

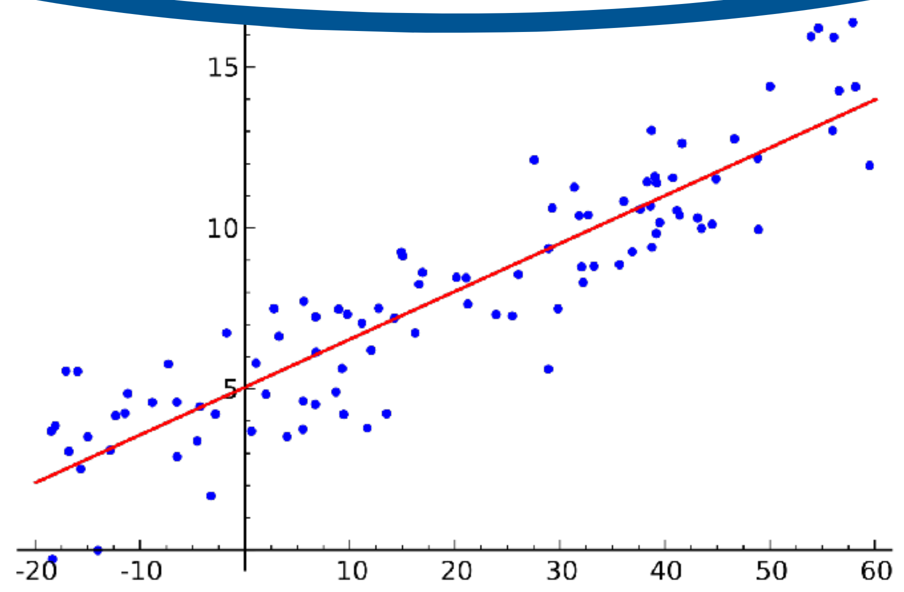
Balamurugan & Bach `16

$\mathcal{X} = \mathcal{Y} =$ simplex
Matrix games / LP

$\mathcal{X} =$ Euclidean, $\mathcal{Y} =$ simplex
Hard margin SVM



$\mathcal{X} = \mathcal{Y} =$ Euclidean
Linear regression



Poster #212

Bilinear games

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top A x, \quad A \in \mathbb{R}^{m \times n}$$

- Simplest case
- Local model for smooth zero-sum game
- Important by themselves

GEOMETRY MATTERS

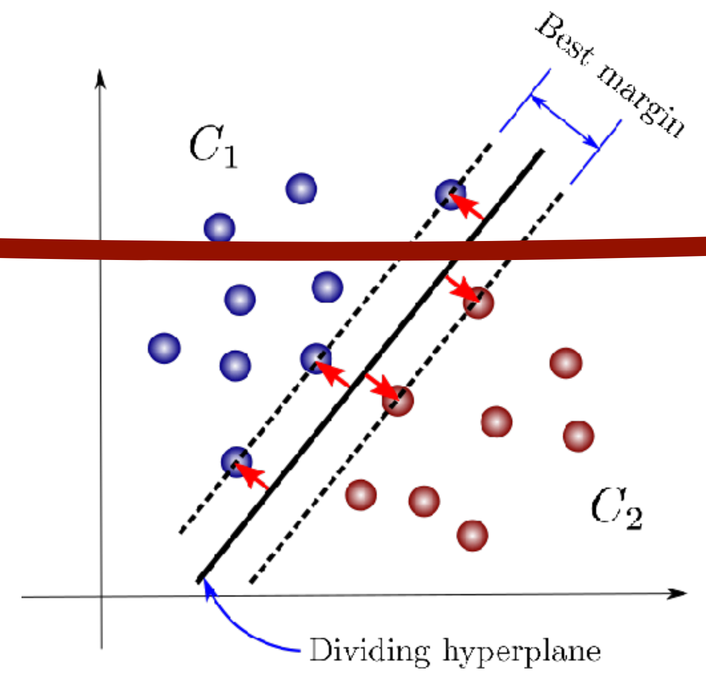
Balamurugan & Bach `16

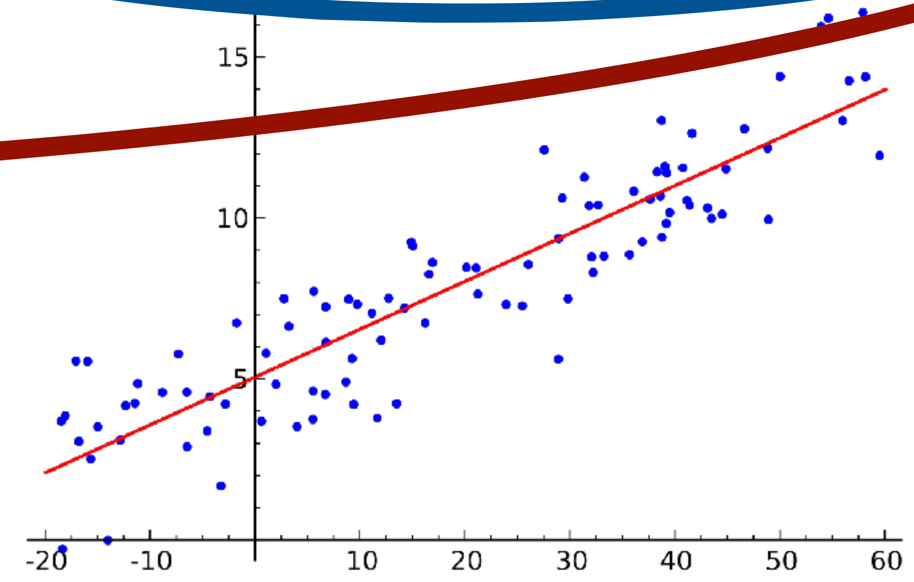
$\mathcal{X} = \mathcal{Y} =$ simplex
Matrix games / LP

$\mathcal{X} =$ Euclidean, $\mathcal{Y} =$ simplex
Hard margin SVM

$\mathcal{X} = \mathcal{Y} =$ Euclidean
Linear regression

Our work





Algorithms and rates

$A \in \mathbb{R}^{m \times n}$

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top A x$$

Algorithms and rates

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top A x$$

$A \in \mathbb{R}^{m \times n}$

for simplicity

$m \asymp n$

$x \mapsto Ax$ takes n^2 time

Algorithms and rates

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top A x$$

$A \in \mathbb{R}^{m \times n}$

for simplicity

$m \asymp n$

$x \mapsto Ax$ takes n^2 time

Exact gradient

(Nemirovski `04, Nesterov `07)

$$n^2 \cdot \frac{L}{\epsilon}$$

Poster #212

Algorithms and rates

$$L = \begin{cases} \max_{ij} |A_{ij}| & \text{simplex-simplex} \\ \max_i \|A_{i:}\|_2 & \text{simplex-ball} \end{cases}$$

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top A x$$

$A \in \mathbb{R}^{m \times n}$

$m \asymp n$

$x \mapsto Ax$ takes n^2 time

for simplicity

Exact gradient

(Nemirovski `04, Nesterov `07)

$$n^2 \cdot \frac{L}{\epsilon}$$

Poster #212

Algorithms and rates

$$L = \begin{cases} \max_{ij} |A_{ij}| & \text{simplex-simplex} \\ \max_i \|A_{i:}\|_2 & \text{simplex-ball} \end{cases}$$

GEOMETRY MATTERS

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top A x$$

$A \in \mathbb{R}^{m \times n}$

$m \asymp n$

$x \mapsto Ax$ takes n^2 time

for simplicity

Exact gradient

(Nemirovski `04, Nesterov `07)

$$n^2 \cdot \frac{L}{\epsilon}$$

Poster #212

Algorithms and rates

$$L = \begin{cases} \max_{ij} |A_{ij}| & \text{simplex-simplex} \\ \max_i \|A_{i:}\|_2 & \text{simplex-ball} \end{cases}$$

GEOMETRY MATTERS

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top A x$$

$A \in \mathbb{R}^{m \times n}$

$m \asymp n$

$x \mapsto Ax$ takes n^2 time

for simplicity

Exact gradient

(Nemirovski `04, Nesterov `07)

$$n^2 \cdot \frac{L}{\epsilon}$$

Stochastic gradient

(GK95, NJLS09, CHW10)

$$n \cdot \frac{L^2}{\epsilon^2}$$

Algorithms and rates

$$L = \begin{cases} \max_{ij} |A_{ij}| & \text{simplex-simplex} \\ \max_i \|A_{i:}\|_2 & \text{simplex-ball} \end{cases}$$

GEOMETRY MATTERS

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} y^\top A x$$

$A \in \mathbb{R}^{m\times n}$

$m \asymp n$

$x \mapsto Ax$ takes n^2 time

for simplicity

Exact gradient

(Nemirovski `04, Nesterov `07)

$$n^2 \cdot \frac{L}{\epsilon}$$

Variance reduction

(our approach)

$$n^2 + n^{3/2} \cdot \frac{L}{\epsilon}$$

Stochastic gradient

(GK95, NJLS09, CHW10)

$$n \cdot \frac{L^2}{\epsilon^2}$$

Poster #212

Algorithms and rates

$$L = \begin{cases} \max_{ij} |A_{ij}| & \text{simplex-simplex} \\ \max_i \|A_{i:}\|_2 & \text{simplex-ball} \end{cases}$$

GEOMETRY MATTERS

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top A x$$

$A \in \mathbb{R}^{m \times n}$

$m \asymp n$ for simplicity

$x \mapsto Ax$ takes n^2 time

Exact gradient

(Nemirovski `04, Nesterov `07)

$$n^2 \cdot \frac{L}{\epsilon}$$

Variance reduction

(our approach)

$$n^2 + n^{3/2} \cdot \frac{L}{\epsilon}$$

Stochastic gradient

(GK95, NJLS09, CHW10)

$$n \cdot \frac{L^2}{\epsilon^2}$$

Image credit: Chawit Waewsawangwong

Poster #212

Algorithms and rates

$$L = \begin{cases} \max_{ij} |A_{ij}| & \text{simplex-simplex} \\ \max_i \|A_{i:}\|_2 & \text{simplex-ball} \end{cases}$$

GEOMETRY MATTERS

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top A x$$

$A \in \mathbb{R}^{m \times n}$

$m \asymp n$

$x \mapsto Ax$ takes n^2 time

for simplicity

Exact gradient

(Nemirovski `04, Nesterov `07)

$$n^2 \cdot \frac{L}{\epsilon}$$

Variance reduction

(our approach)

$$n^2 + n^{3/2} \cdot \frac{L}{\epsilon}$$

Stochastic gradient

(GK95, NJLS09, CHW10)

$$n \cdot \frac{L^2}{\epsilon^2}$$

Image credit: Chawit Waewsawangwong

Algorithms and rates

$$L = \begin{cases} \max_{ij} |A_{ij}| & \text{simplex-simplex} \\ \max_i \|A_{i:}\|_2 & \text{simplex-ball} \end{cases}$$

GEOMETRY MATTERS

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} y^\top A x$$

$A \in \mathbb{R}^{m \times n}$

$m \asymp n$

$x \mapsto Ax$ takes n^2 time

for simplicity

Exact gradient

(Nemirovski `04, Nesterov `07)

$$n^2 \cdot \frac{L}{\epsilon}$$

Variance reduction

(our approach)

$$n^2 + n^{3/2} \cdot \frac{L}{\epsilon}$$

Stochastic gradient

(GK95, NJLS09, CHW10)

$$n \cdot \frac{L^2}{\epsilon^2}$$

Image credit: Chawit Waewsawangwong

Poster #212

Algorithms and rates

$$L = \begin{cases} \max_{ij} |A_{ij}| & \text{simplex-simplex} \\ \max_i \|A_{i:}\|_2 & \text{simplex-ball} \end{cases}$$

GEOMETRY MATTERS

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top A x$$

$A \in \mathbb{R}^{m \times n}$

$m \asymp n$

$x \mapsto Ax$ takes n^2 time

for simplicity

Exact gradient

(Nemirovski `04, Nesterov `07)

$$n^2 \cdot \frac{L}{\epsilon}$$

VR always better

GEOMETRY MATTERS

Variance reduction

(our approach)

$$n^2 + n^{3/2} \cdot \frac{L}{\epsilon}$$

Stochastic gradient

(GK95, NJLS09, CHW10)

$$n \cdot \frac{L^2}{\epsilon^2}$$

Image credit: Chawit Waewsawangwong

Poster #212

Algorithms and rates

$$L = \begin{cases} \max_{ij} |A_{ij}| & \text{simplex-simplex} \\ \max_i \|A_{i:}\|_2 & \text{simplex-ball} \end{cases}$$

GEOMETRY MATTERS

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} y^\top A x$$

$A \in \mathbb{R}^{m\times n}$

$m \asymp n$

$x \mapsto Ax$ takes n^2 time

for simplicity

Exact gradient

(Nemirovski `04, Nesterov `07)

$$n^2 \cdot \frac{L}{\epsilon}$$

VR always better

GEOMETRY MATTERS

Variance reduction

(our approach)

$$n^2 + n^{3/2} \cdot \frac{L}{\epsilon}$$

VR better for $\Omega(1)$ passes over data

Stochastic gradient

(GK95, NJLS09, CHW10)

$$n \cdot \frac{L^2}{\epsilon^2}$$

Image credit: Chawit Waewsawangwong

Poster #212

It's all in the gradient estimator

Reference point

x_0

Poster #212

It's all in the gradient estimator

Centered gradient estimator $g_{x_0}(\cdot)$

$$\mathbb{E}\, g_{x_0}(x) = \nabla f(x) \text{ and } \mathbb{E}\|g_{x_0}(x) - \nabla f(x)\|_*^2 \le L^2\|x - x_0\|^2$$

Reference point

x_0

Also using this concept in the Euclidean setting: VR for non-convex optimization (AH`16, RHSPS`16, FLLZ`18, ZXG`18) & bilinear saddle-point problems (BB`16)

It's all in the gradient estimator

Centered gradient estimator $g_{x_0}(\cdot)$

$$\mathbb{E}\, g_{x_0}(x) = \nabla f(x) \text{ and } \underbrace{\mathbb{E}\|g_{x_0}(x) - \nabla f(x)\|_*^2}_{\text{variance}} \le L^2\|x - x_0\|^2$$

Reference point $\longrightarrow \bullet\, x_0$

Also using this concept in the Euclidean setting: VR for non-convex optimization (AH`16, RHSPS`16, FLLZ`18, ZXG`18) & bilinear saddle-point problems (BB`16)

Poster #212

It's all in the gradient estimator

Centered gradient estimator $g_{x_0}(\,\cdot\,)$

$$\mathbb{E}\, g_{x_0}(x) = \nabla f(x) \text{ and } \mathbb{E}\|g_{x_0}(x) - \nabla f(x)\|_*^2 \leq L^2\|x - x_0\|^2$$

variance

distance from reference point

Reference point

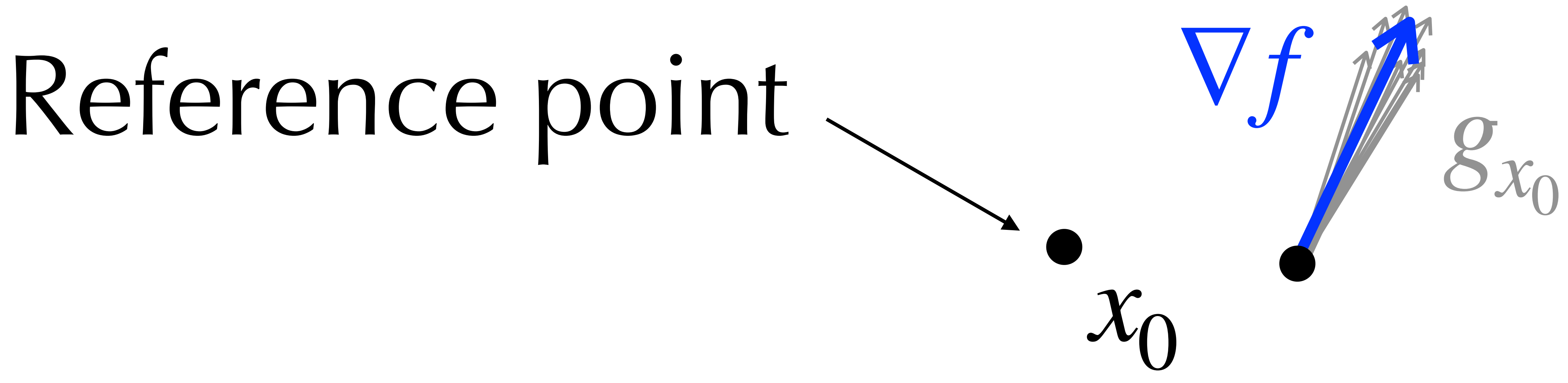
x_0

Also using this concept in the Euclidean setting: VR for non-convex optimization (AH`16, RHSPS`16, FLLZ`18, ZXG`18) & bilinear saddle-point problems (BB`16)

It's all in the gradient estimator

Centered gradient estimator $g_{x_0}(\cdot)$

$$\mathbb{E}\, g_{x_0}(x) = \nabla f(x) \text{ and } \underbrace{\mathbb{E}\|g_{x_0}(x) - \nabla f(x)\|_*^2}_{\text{variance}} \leq L^2 \underbrace{\|x - x_0\|^2}_{\text{distance from reference point}}$$



Also using this concept in the Euclidean setting: VR for non-convex optimization (AH`16, RHSPS`16, FLLZ`18, ZXG`18) & bilinear saddle-point problems (BB`16)

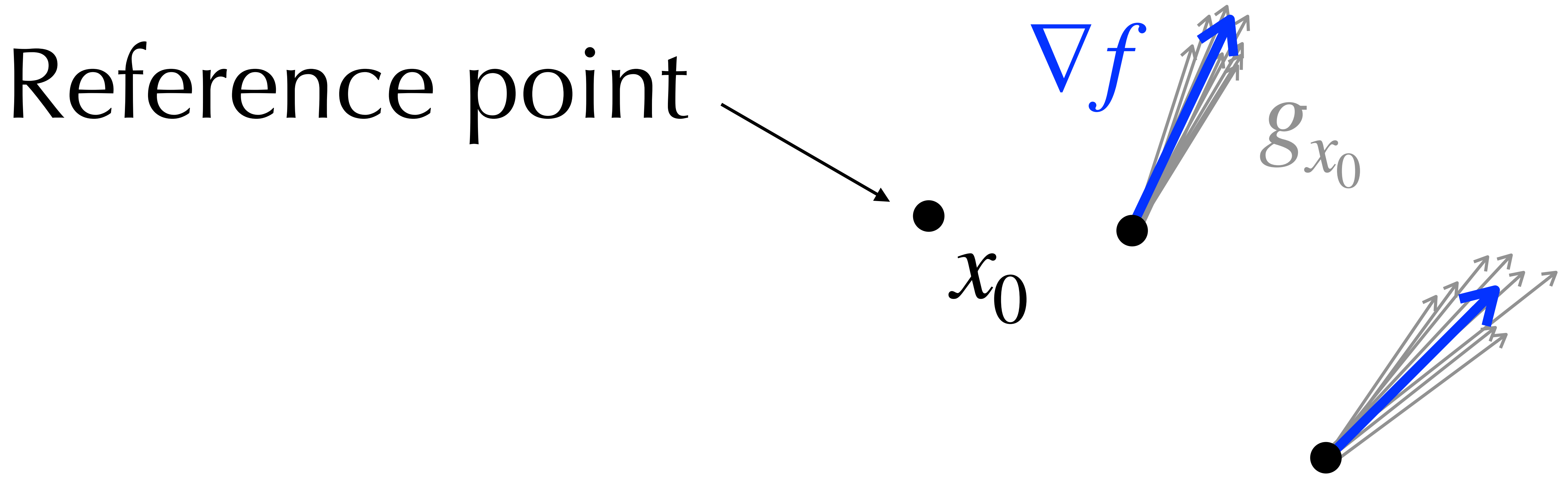
It's all in the gradient estimator

Centered gradient estimator $g_{x_0}(\,\cdot\,)$

$$\mathbb{E}\, g_{x_0}(x) = \nabla f(x) \text{ and } \mathbb{E}\|g_{x_0}(x) - \nabla f(x)\|_*^2 \le L^2\|x - x_0\|^2$$

variance

distance from reference point



Also using this concept in the Euclidean setting: VR for non-convex optimization (AH`16, RHSPS`16, FLLZ`18, ZXG`18) & bilinear saddle-point problems (BB`16)

It's all in the gradient estimator

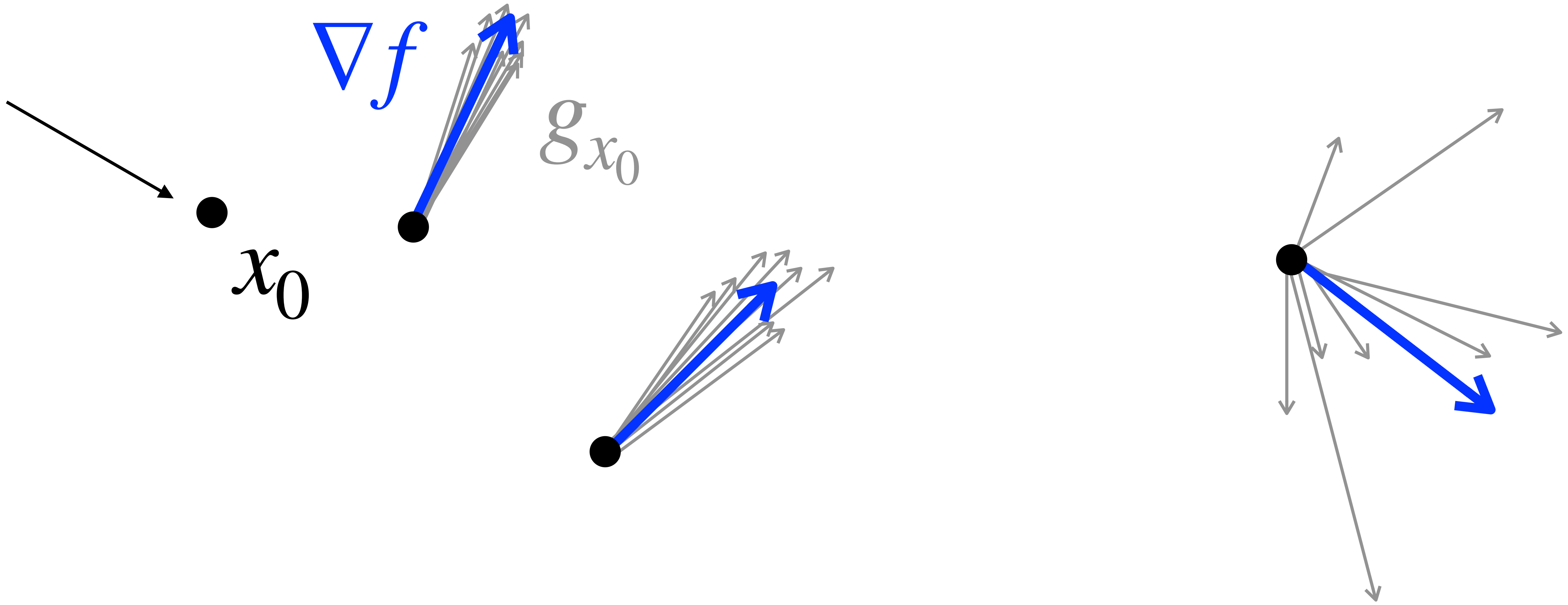
Centered gradient estimator $g_{x_0}(\cdot)$

$$\mathbb{E}\, g_{x_0}(x) = \nabla f(x) \text{ and } \mathbb{E}\|g_{x_0}(x) - \nabla f(x)\|_*^2 \le L^2\|x - x_0\|^2$$

variance

distance from reference point

Reference point



Also using this concept in the Euclidean setting: VR for non-convex optimization (AH`16, RHSPS`16, FLLZ`18, ZXG`18) & bilinear saddle-point problems (BB`16)

Poster #212

It's all in the gradient estimator

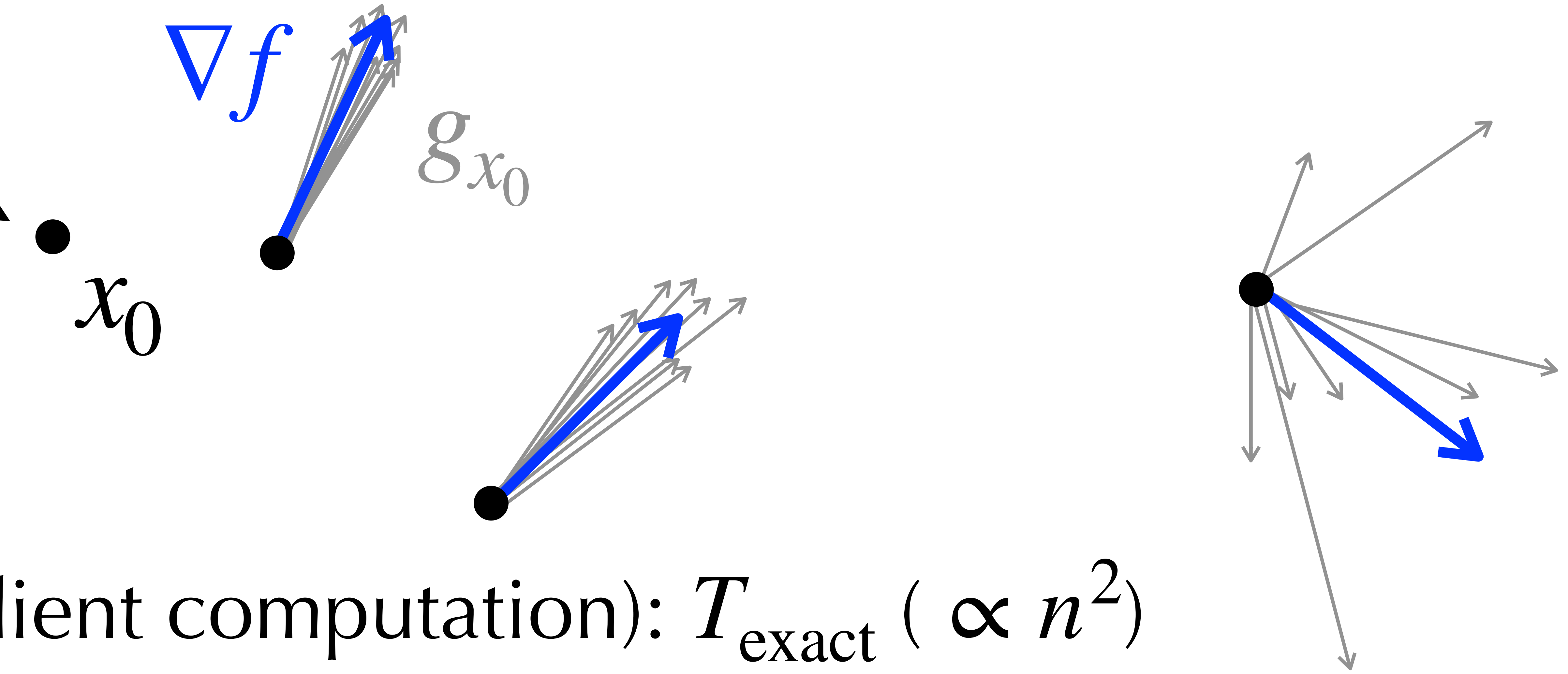
Centered gradient estimator $g_{x_0}(\,\cdot\,)$

$$\mathbb{E}\, g_{x_0}(x) = \nabla f(x) \text{ and } \mathbb{E}\|g_{x_0}(x) - \nabla f(x)\|_*^2 \leq L^2\|x - x_0\|^2$$

variance

distance from reference point

Reference point



Typical cost:

- Preprocessing (exact gradient computation): T_{exact} ($\propto n^2$)
- Per stochastic gradient: T_{stoch} ($\propto n$)

Also using this concept in the Euclidean setting: VR for non-convex optimization (AH`16, RHSPS`16, FLLZ`18, ZXG`18) & bilinear saddle-point problems (BB`16)

Constructing a centered estimator

$$\min_{x \in \text{simplex}} \max_{y \in \text{simplex}} y^\top A x$$

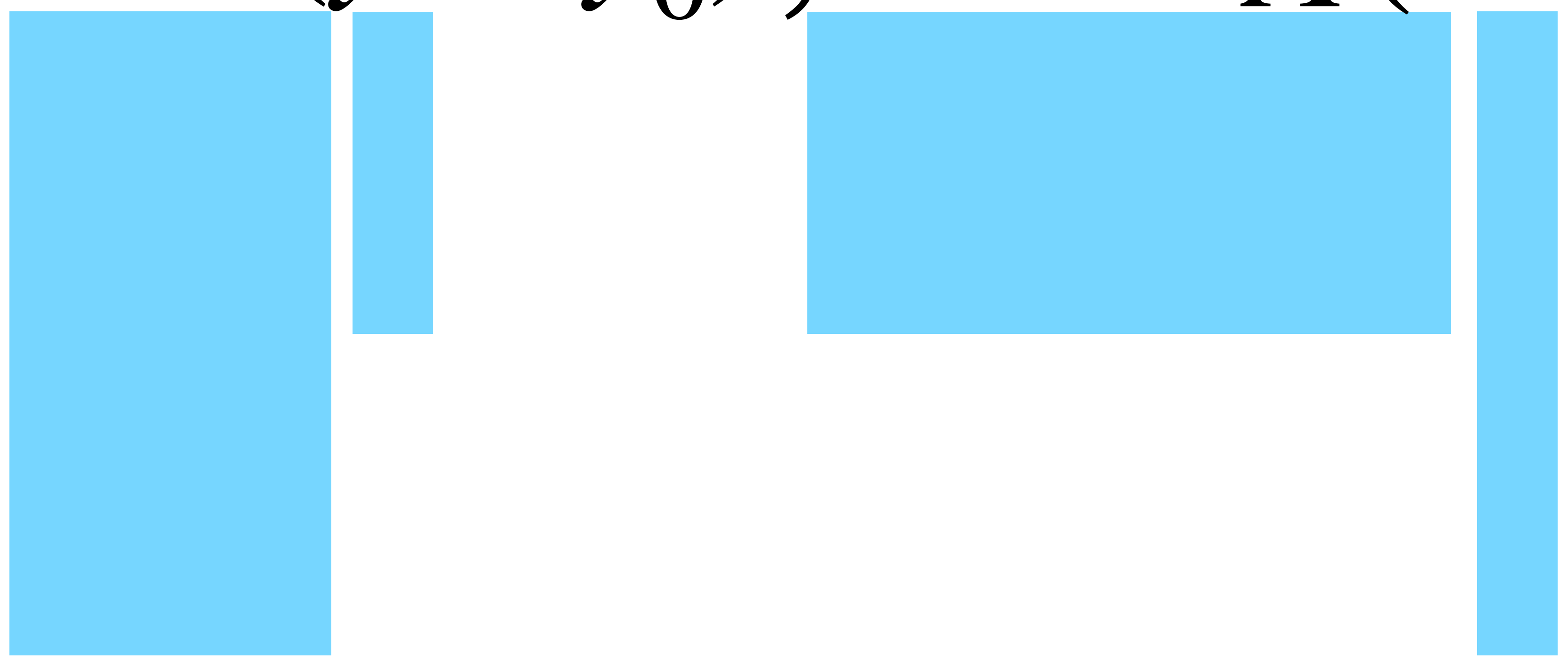
$$\nabla f(x, y) = [A^\top y, Ax]$$

Constructing a centered estimator

$$\min_{x \in \text{simplex}} \max_{y \in \text{simplex}} y^\top A x$$

gradient at reference point

$$\nabla f(x, y) = \underbrace{[A^\top y_0, A x_0]}^{\text{gradient at reference point}} + \left[A^\top (y - y_0),\ A(x - x_0) \right]$$



Constructing a centered estimator

$$\min_{x \in \text{simplex}} \max_{y \in \text{simplex}} y^\top A x$$

gradient at reference point

$$\nabla f(x,y) = [A^\top y_0, A x_0] + \left[A^\top (y - y_0), \quad A(x - x_0) \right]$$

$$g_{x_0,y_0}(x,y) = [A^\top y_0, A x_0] + \left[\quad , \quad \right]$$

$T_{\text{exact}} = O(n^2)$

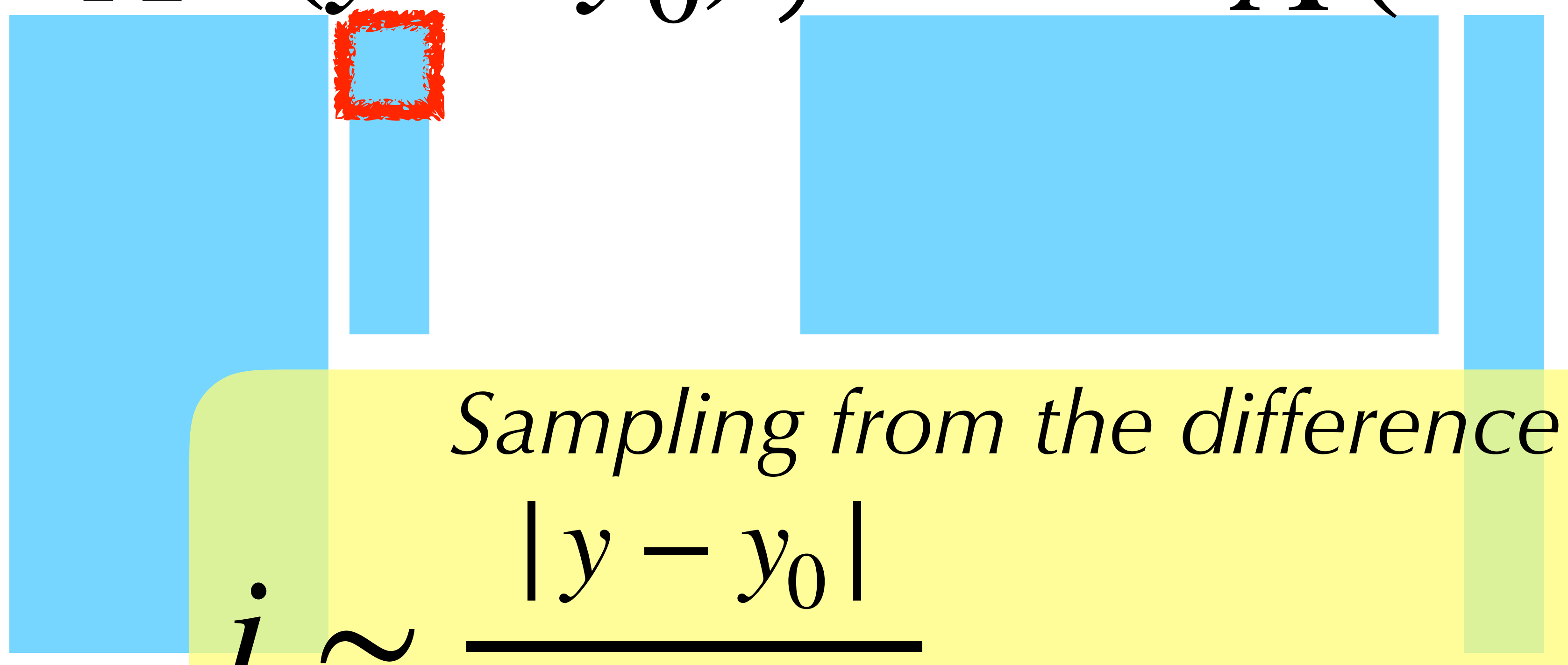
preprocessing cost

Constructing a centered estimator

$$\min_{x \in \text{simplex}} \max_{y \in \text{simplex}} y^\top A x$$

gradient at reference point

$$\nabla f(x,y) = [A^\top y_0, A x_0] + \left[A^\top (y - y_0), \quad A(x - x_0) \right]$$



Sampling from the difference

$$i \sim \frac{|y - y_0|}{\|y - y_0\|_1}$$

$$g_{x_0,y_0}(x,y) = [A^\top y_0, A x_0] + [\quad , \quad]$$

$$\uparrow$$

$$T_{\text{exact}} = O(n^2)$$

preprocessing cost

Poster #212

Constructing a centered estimator

$$\min_{x\in\text{simplex}} \; \max_{y\in\text{simplex}} \; y^\top A x$$

gradient at reference point

$$\nabla f(x, y) = [A^\top y_0, Ax_0] + \left[A^\top(y - y_0), \quad A(x - x_0) \right]$$

Sampling from the difference

$$i \sim \frac{|y - y_0|}{\|y - y_0\|_1}$$

$$g_{x_0,y_0}(x, y) = [A^\top y_0, Ax_0] + \left[\frac{\|y - y_0\|_1}{\text{sign}([y - y_0]_i)} A_{i:} \, , \quad \right]$$

$$\uparrow$$

$T_{\text{exact}} = O(n^2)$

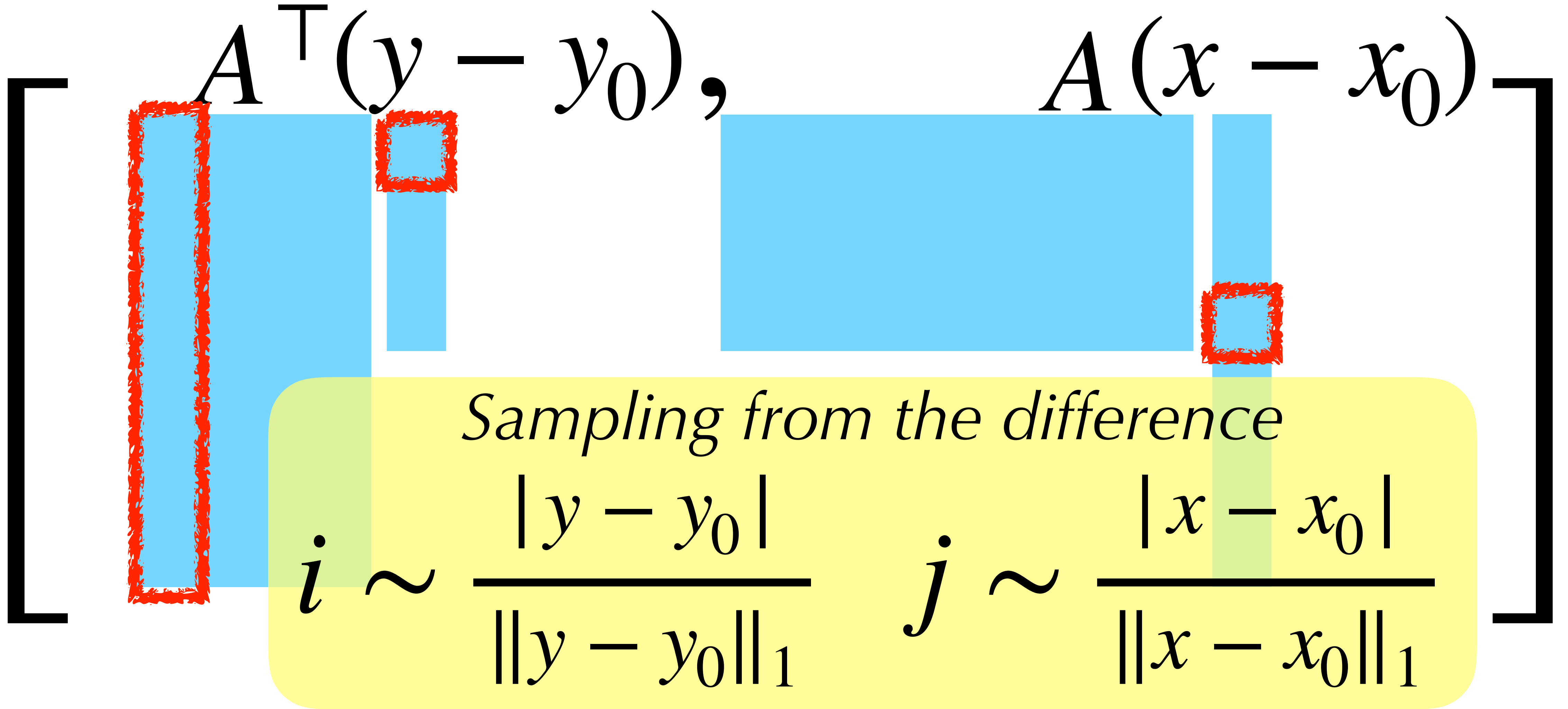
preprocessing cost

Constructing a centered estimator

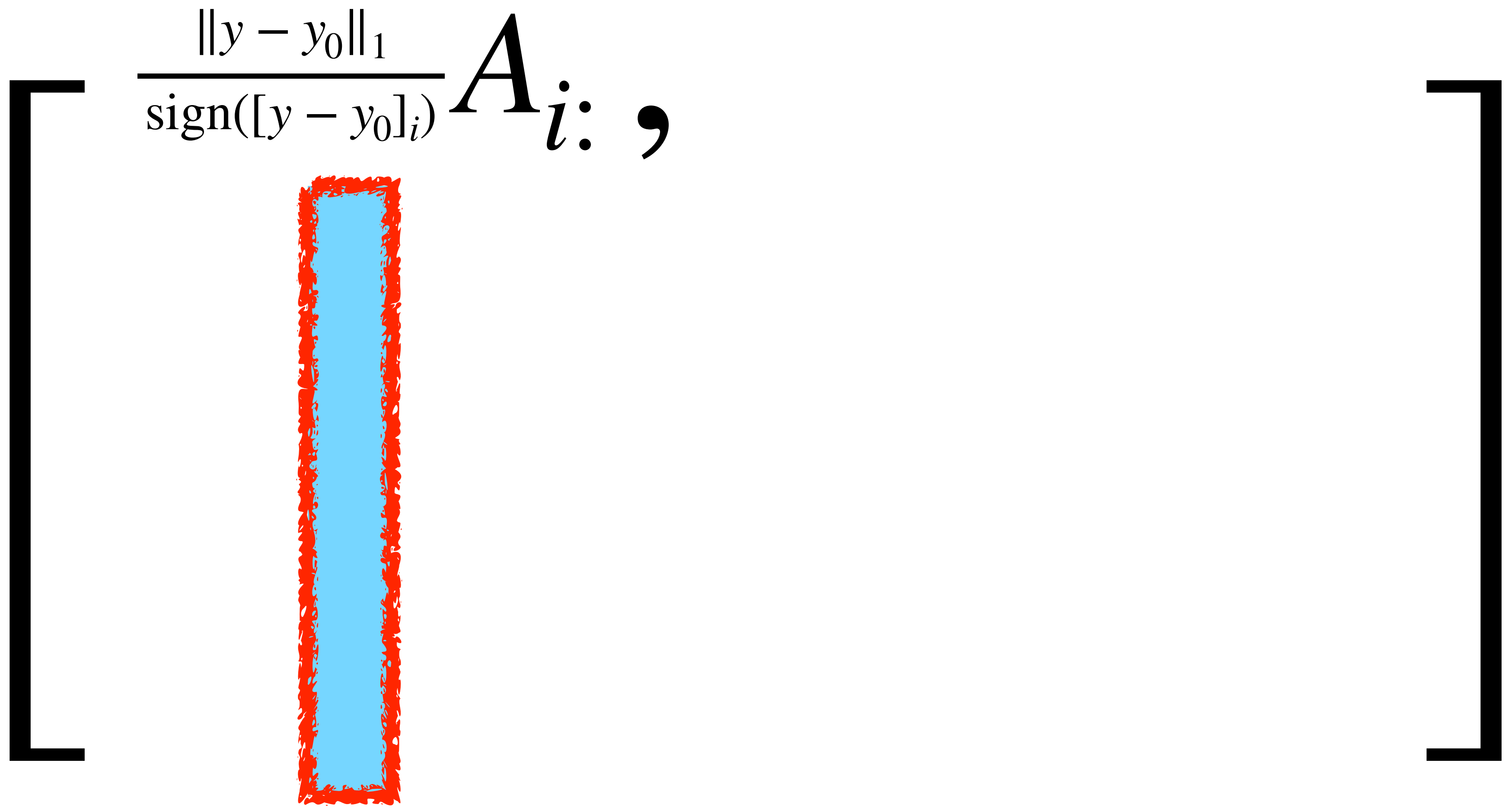
$$\min_{x \in \text{simplex}} \max_{y \in \text{simplex}} y^\top A x$$

gradient at reference point

$$\nabla f(x,y) = [A^\top y_0, A x_0] +$$



$$g_{x_0, y_0}(x,y) = [A^\top y_0, A x_0] +$$



$T_{\text{exact}} = O(n^2)$

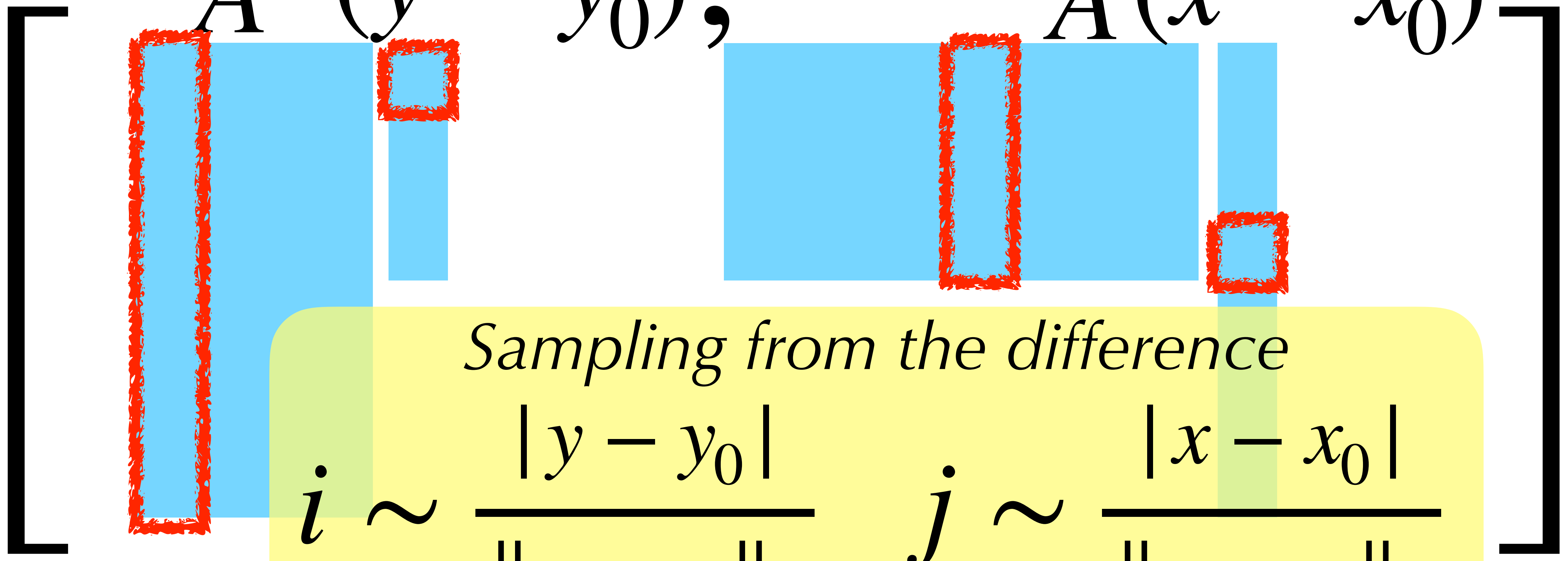
preprocessing cost

Constructing a centered estimator

$$\min_{x \in \text{simplex}} \max_{y \in \text{simplex}} y^\top A x$$

gradient at reference point

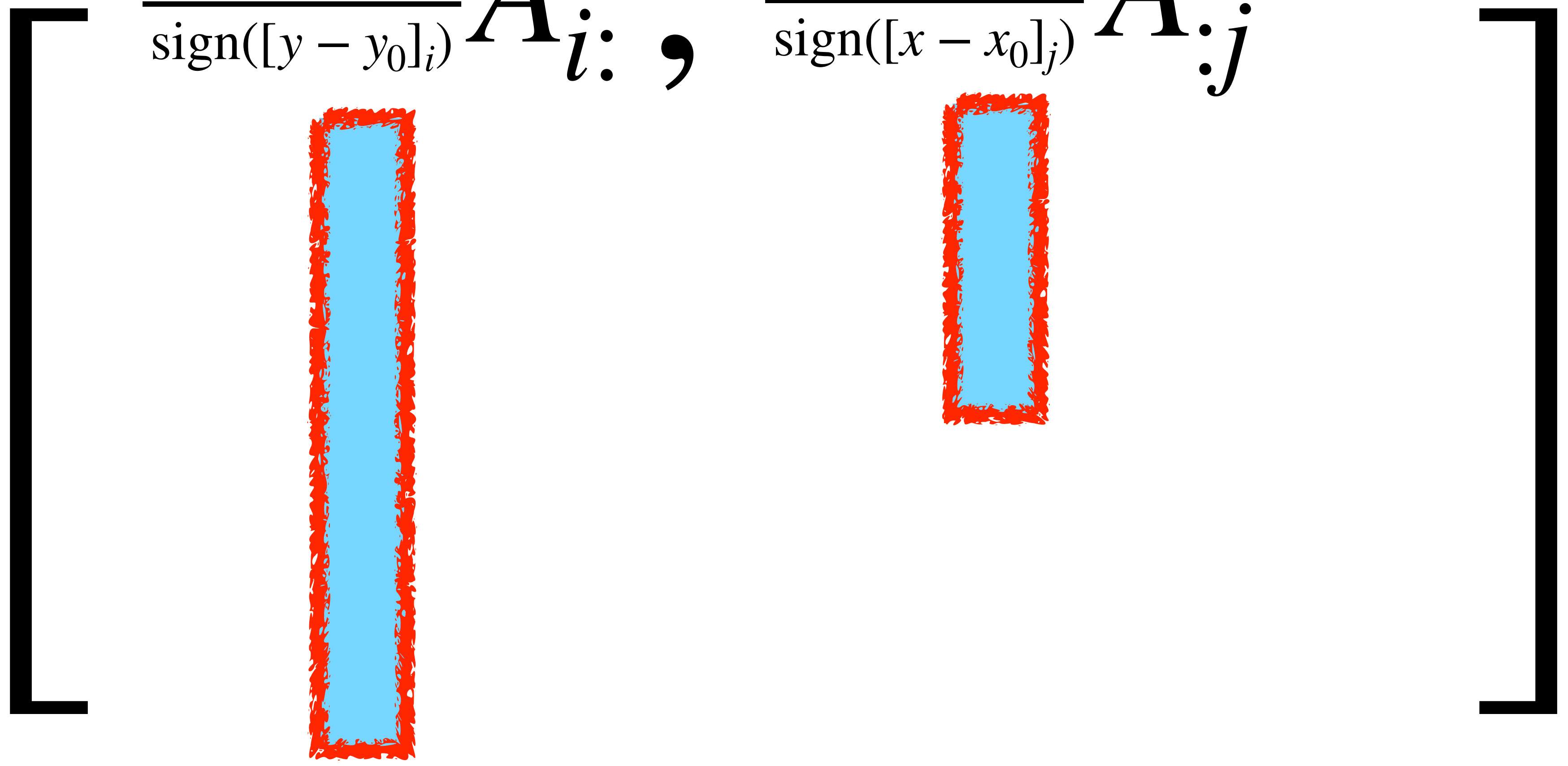
$$\nabla f(x, y) = [A^\top y_0, A x_0] + \left[A^\top (y - y_0),\ A(x - x_0) \right]$$



Sampling from the difference

$$i \sim \frac{|y - y_0|}{\|y - y_0\|_1} \qquad j \sim \frac{|x - x_0|}{\|x - x_0\|_1}$$

$$g_{x_0, y_0}(x, y) = [A^\top y_0, A x_0] + \left[\frac{\|y - y_0\|_1}{\operatorname{sign}([y - y_0]_i)} A_{i:},\ \frac{\|x - x_0\|_1}{\operatorname{sign}([x - x_0]_j)} A_{:j} \right]$$



$$T_{\text{exact}} = O(n^2)$$

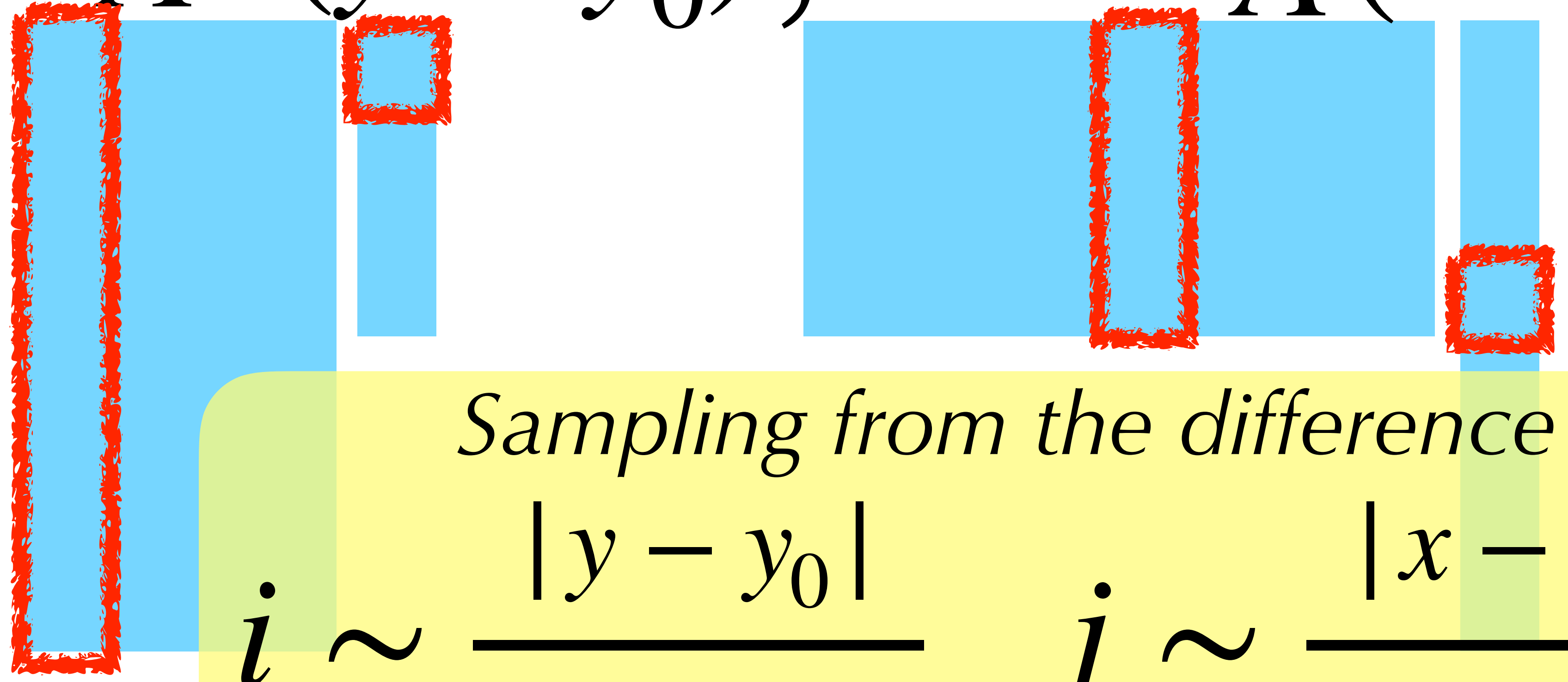
preprocessing cost

Constructing a centered estimator

$$\min_{x \in \text{simplex}} \max_{y \in \text{simplex}} y^\top A x$$

gradient at reference point

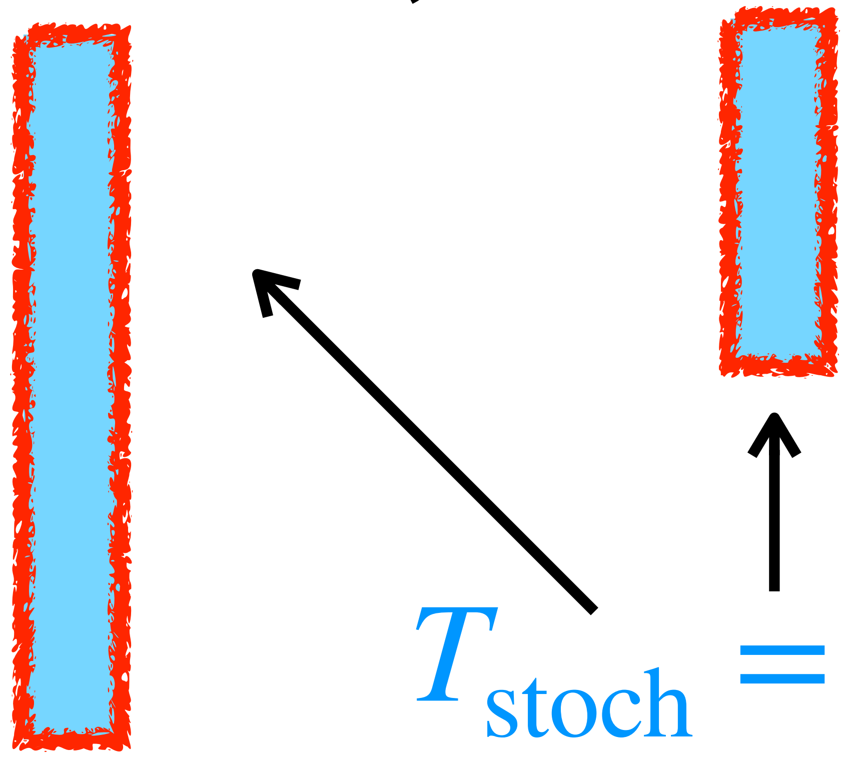
$$\nabla f(x,y) = [A^\top y_0, Ax_0] + \left[A^\top (y - y_0), \quad A(x - x_0) \right]$$



Sampling from the difference

$$i \sim \frac{|y - y_0|}{\|y - y_0\|_1} \qquad j \sim \frac{|x - x_0|}{\|x - x_0\|_1}$$

$$g_{x_0,y_0}(x,y) = [A^\top y_0, Ax_0] + \left[\frac{\|y - y_0\|_1}{\operatorname{sign}([y - y_0]_i)} A_{i:}, \quad \frac{\|x - x_0\|_1}{\operatorname{sign}([x - x_0]_j)} A_{:j} \right]$$



$T_{\text{exact}} = O(n^2)$

preprocessing cost

$T_{\text{stoch}} = O(n)$

per-estimation cost

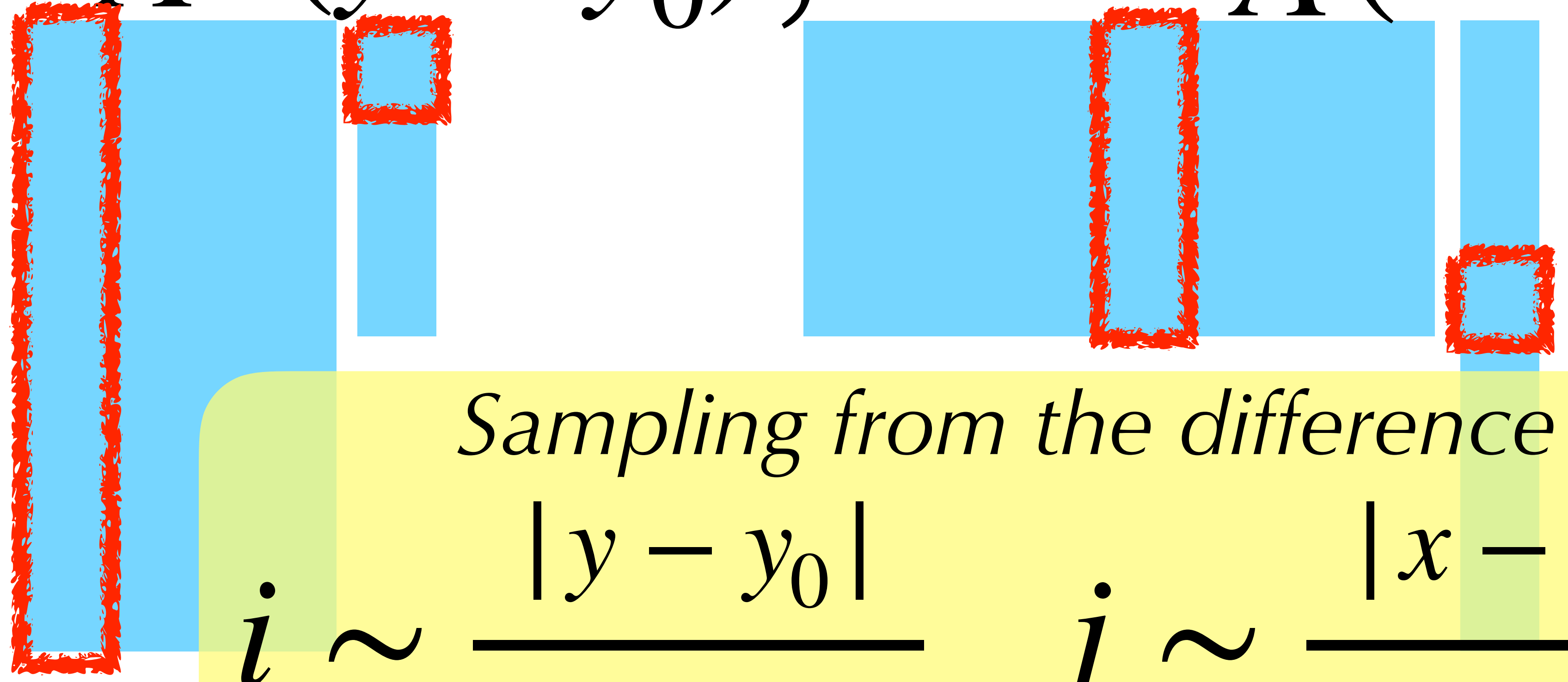
Poster #212

Constructing a centered estimator

$$\min_{x \in \text{simplex}} \max_{y \in \text{simplex}} y^\top A x$$

gradient at reference point

$$\nabla f(x,y) = [A^\top y_0, A x_0] + \left[A^\top(y - y_0), \quad A(x - x_0) \right]$$

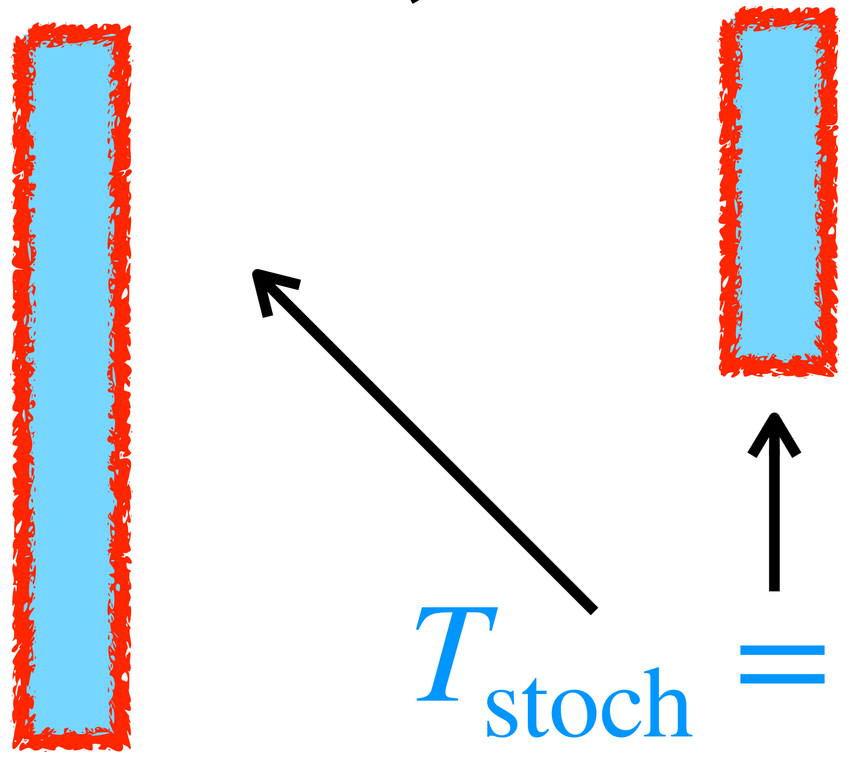


Sampling from the difference

$$i \sim \frac{|y - y_0|}{\|y - y_0\|_1} \qquad j \sim \frac{|x - x_0|}{\|x - x_0\|_1}$$

$$\mathbb{E}\|g_{x_0,y_0}(x,y) - \nabla f(x,y)\|_\infty^2 \leq L^2 \|[x,y] - [x_0,y_0]\|_1^2$$

$$g_{x_0,y_0}(x,y) = [A^\top y_0, A x_0] + \left[\frac{\|y - y_0\|_1}{\text{sign}([y - y_0]_i)} A_{i:}, \quad \frac{\|x - x_0\|_1}{\text{sign}([x - x_0]_j)} A_{:j} \right]$$



$T_{\text{exact}} = O(n^2)$
preprocessing cost

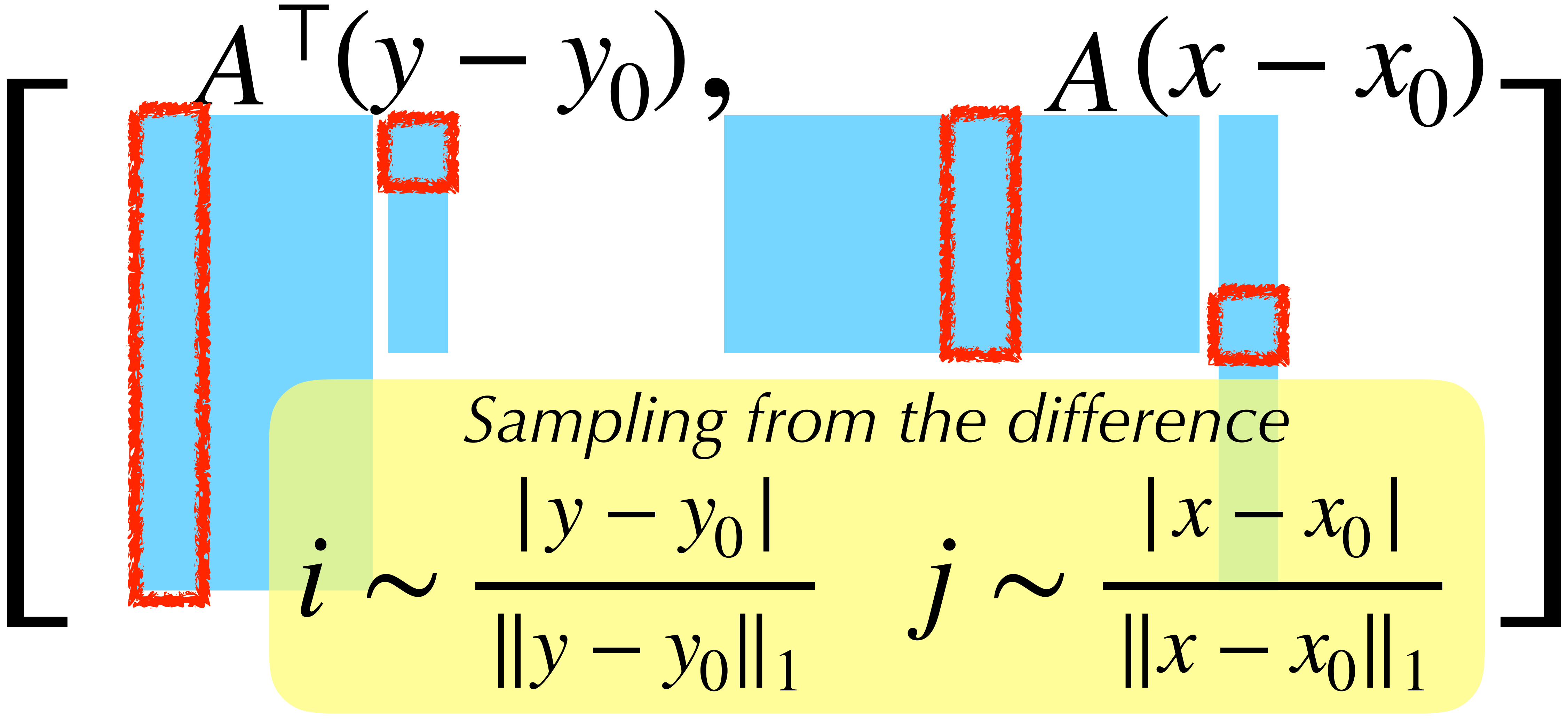
$T_{\text{stoch}} = O(n)$
per-estimation cost

Constructing a centered estimator

$$\min_{x \in \text{simplex}} \max_{y \in \text{simplex}} y^\top A x$$

gradient at reference point

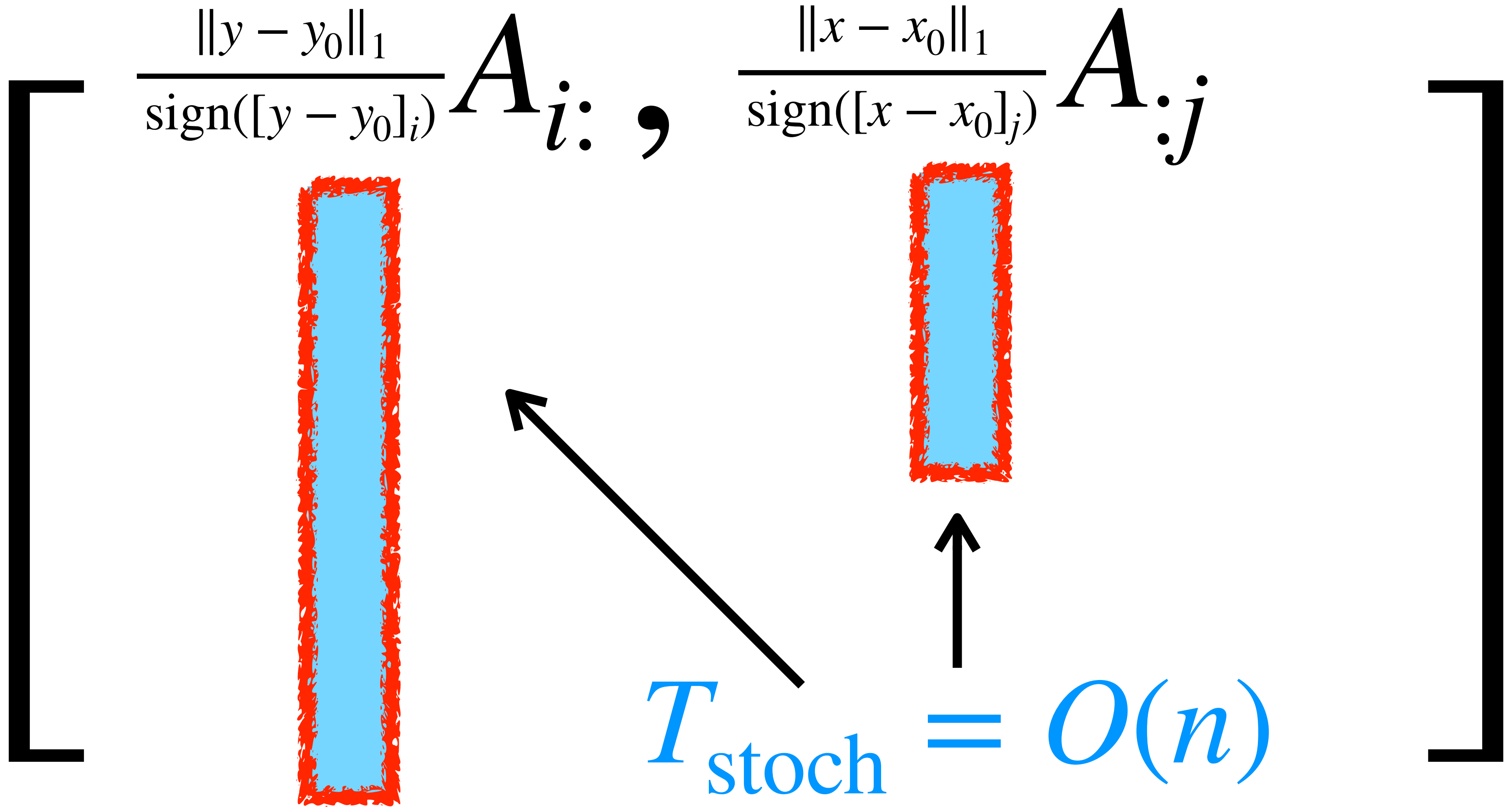
$$\nabla f(x,y) = [A^\top y_0, A x_0] +$$



$$\mathbb{E}\|g_{x_0,y_0}(x,y) - \nabla f(x,y)\|_\infty^2 \le L^2 \|[x,y] - [x_0,y_0]\|_1^2$$

GEOMETRY MATTERS

$$g_{x_0,y_0}(x,y) = [A^\top y_0, A x_0] +$$



$T_{\text{exact}} = O(n^2)$
preprocessing cost

$T_{\text{stoch}} = O(n)$
per-estimation cost

Variance reduction framework

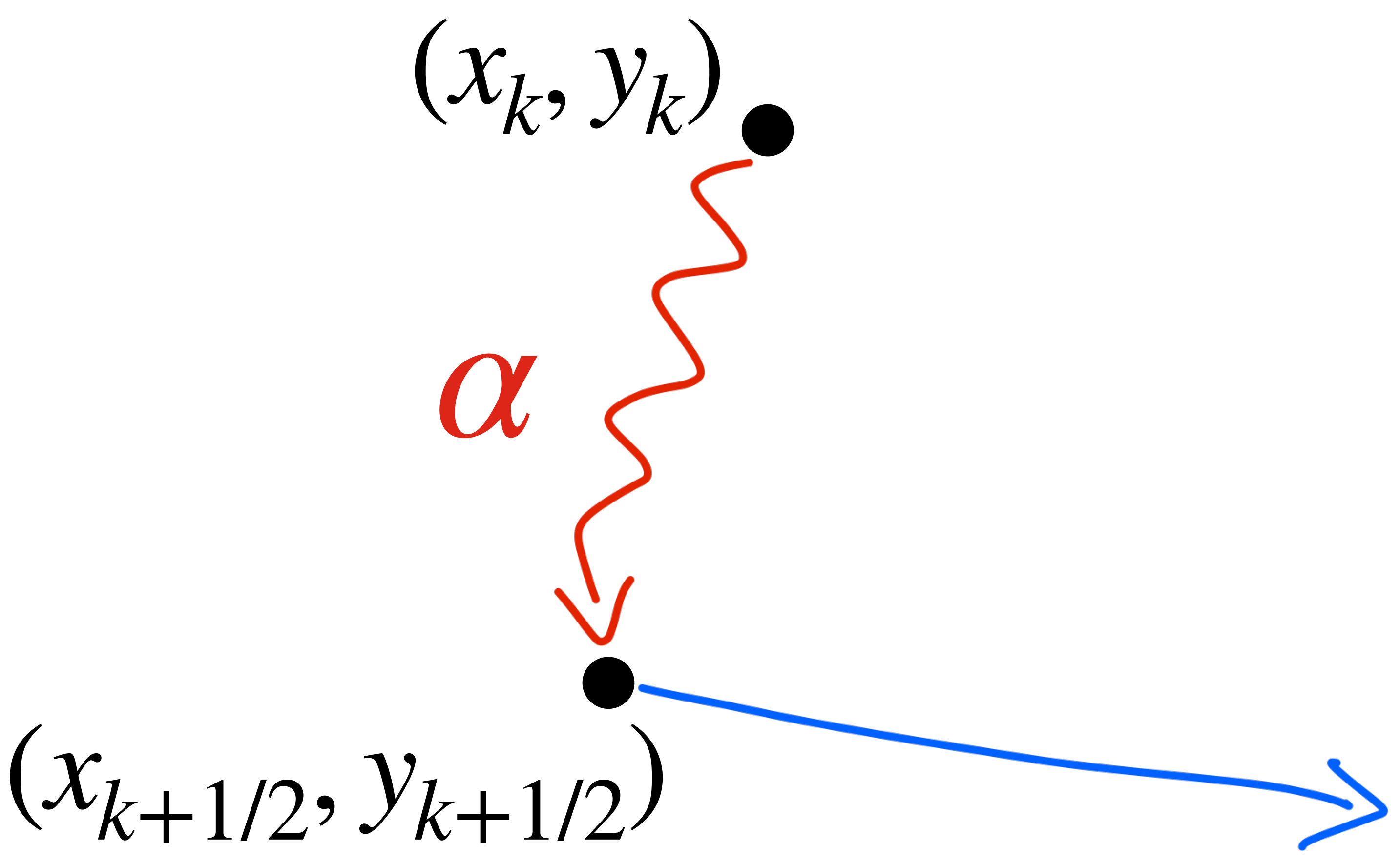
Method	# of iterations	cost per iteration
Basic proximal method (with parameter α)		
$(x_{k+1}, y_{k+1}) \leftarrow \arg\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \left\{ f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2 \right\}$	$\frac{\alpha}{\epsilon}$	cost of prox

Variance reduction framework

Method	# of iterations	cost per iteration
Basic proximal method (with parameter α)		
$(x_{k+1}, y_{k+1}) \leftarrow \arg\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \left\{ f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2 \right\}$	$\frac{\alpha}{\epsilon}$	cost of prox
Nemirovski's "conceptual prox-method"		
$(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to proximal problem $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox + T_{exact}

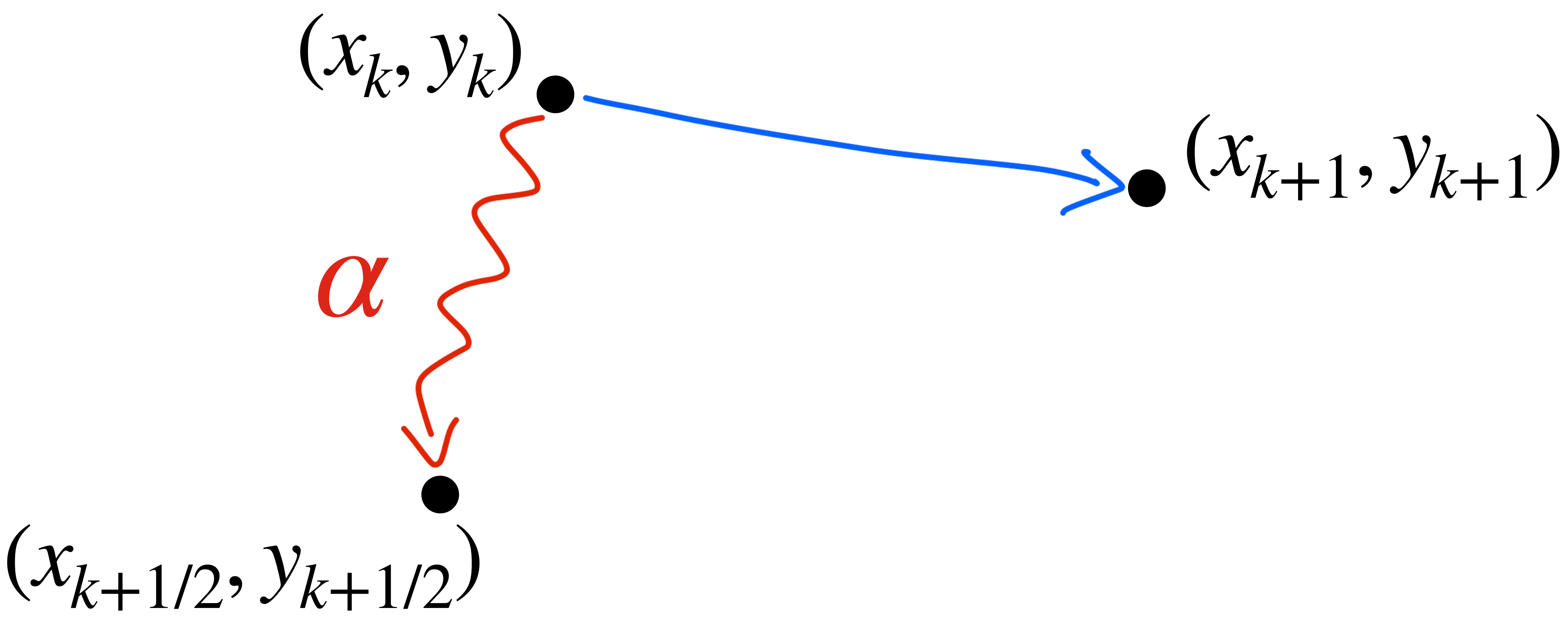
Variance reduction framework

Method	# of iterations	cost per iteration
Basic proximal method (with parameter α) $(x_{k+1}, y_{k+1}) \leftarrow \arg\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} \left\{ f(x,y) + \frac{\alpha}{2}\lVert x - x_k\rVert^2 - \frac{\alpha}{2}\lVert y - y_k\rVert^2 \right\}$	$\frac{\alpha}{\epsilon}$	cost of prox
Nemirovski's "conceptual prox-method" $(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to proximal problem $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox $+ \ T_{\text{exact}}$



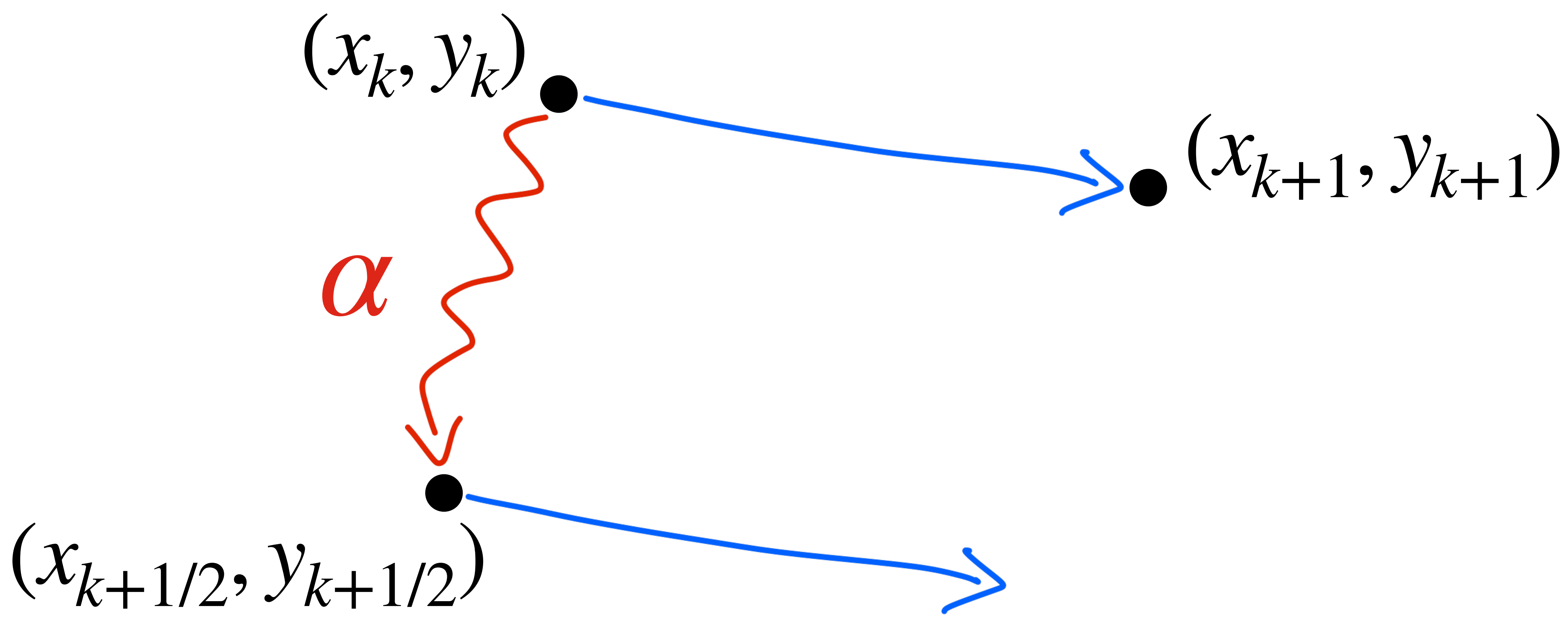
Variance reduction framework

Method	# of iterations	cost per iteration
<u>Basic proximal method (with parameter α)</u> $(x_{k+1}, y_{k+1}) \leftarrow \arg\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} \left\{ f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2 \right\}$	$\frac{\alpha}{\epsilon}$	cost of prox
<u>Nemirovski's "conceptual prox-method"</u> $(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to proximal problem $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox + T_{exact}



Variance reduction framework

Method	# of iterations	cost per iteration
<u>Nemirovski's "conceptual prox-method"</u> $(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to $f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2$ $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox $+\ T_{\text{exact}}$

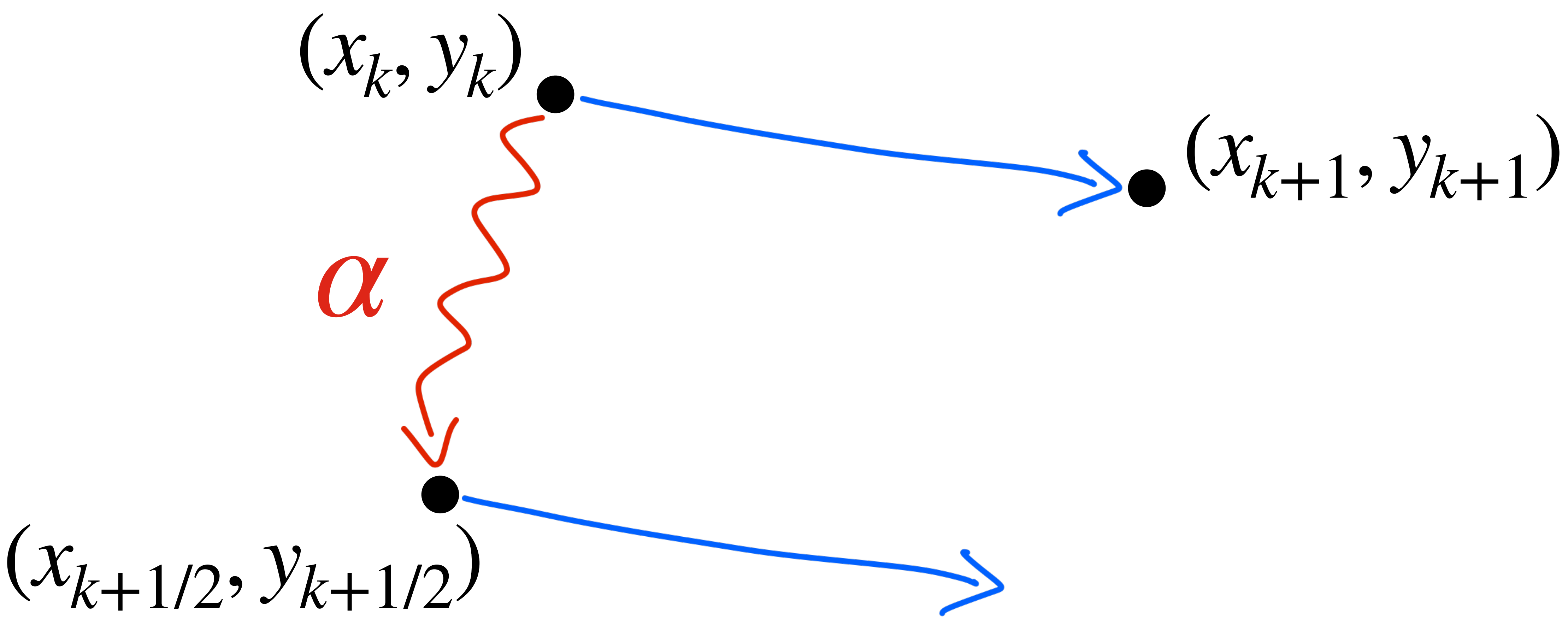


Poster #212

Variance reduction framework

Method	# of iterations	cost per iteration
<u>Nemirovski's "conceptual prox-method"</u> $(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to $f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2$ $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox $+ T_{\text{exact}}$

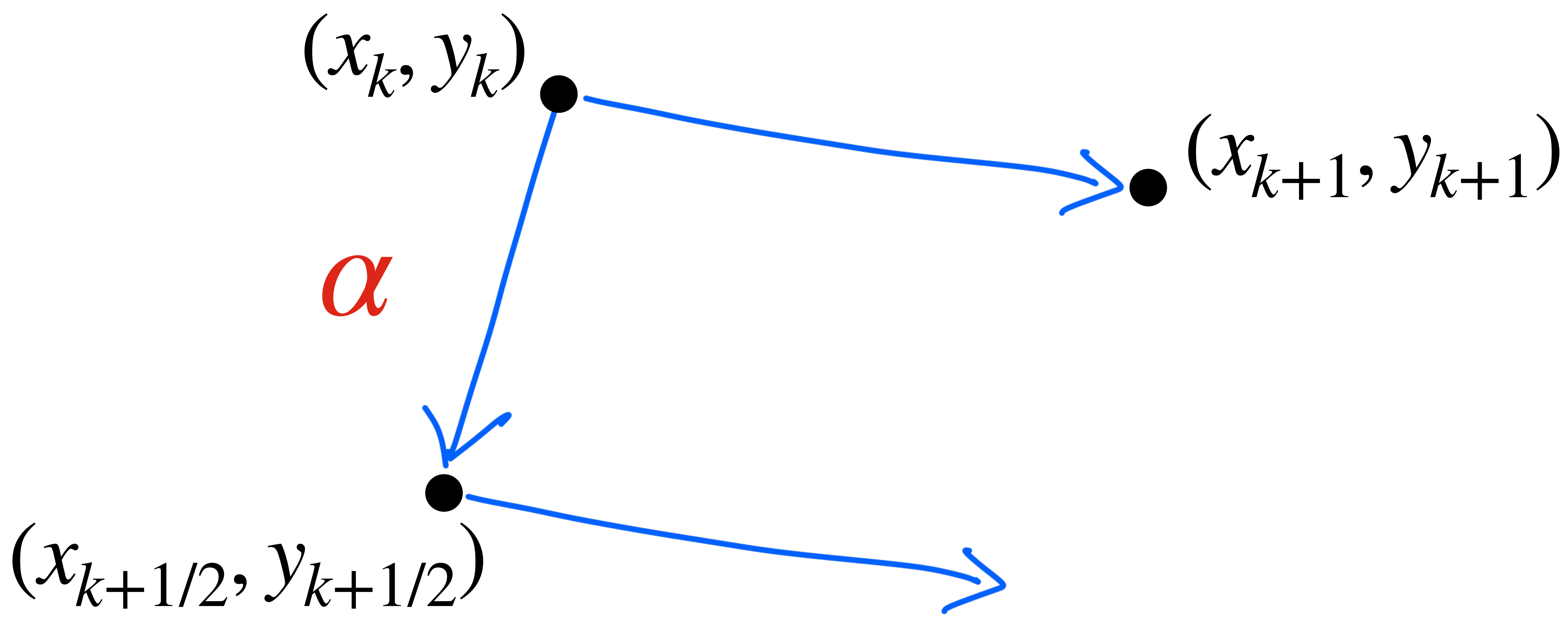
↳ <u>Mirror-prox:</u> rough solution = a gradient step



Variance reduction framework

Method	# of iterations	cost per iteration
Nemirovski's "conceptual prox-method" $(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to $f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2$ $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox $+ T_{\text{exact}}$

↳ Mirror-prox: rough solution = a gradient step

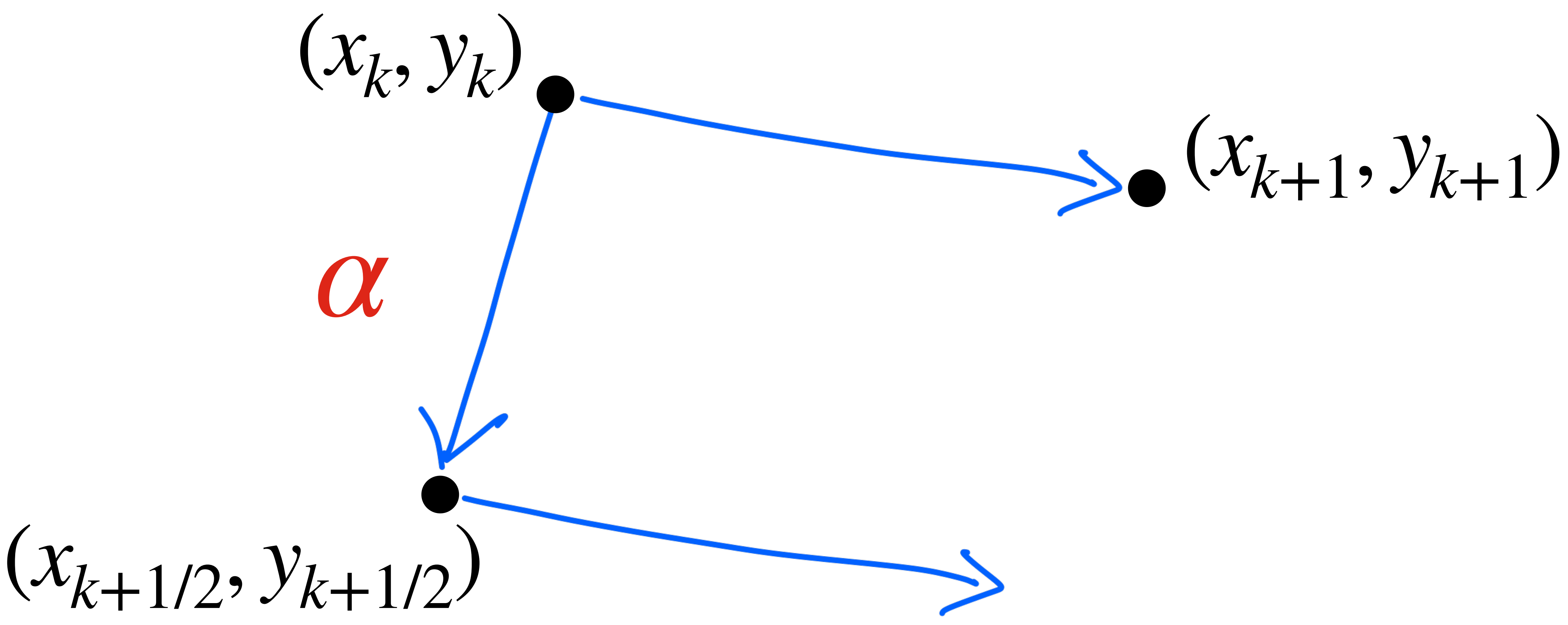


Poster #212

Variance reduction framework

Method	# of iterations	cost per iteration
<u>Nemirovski's "conceptual prox-method"</u> $(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to $f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2$ $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox $+ T_{\text{exact}}$

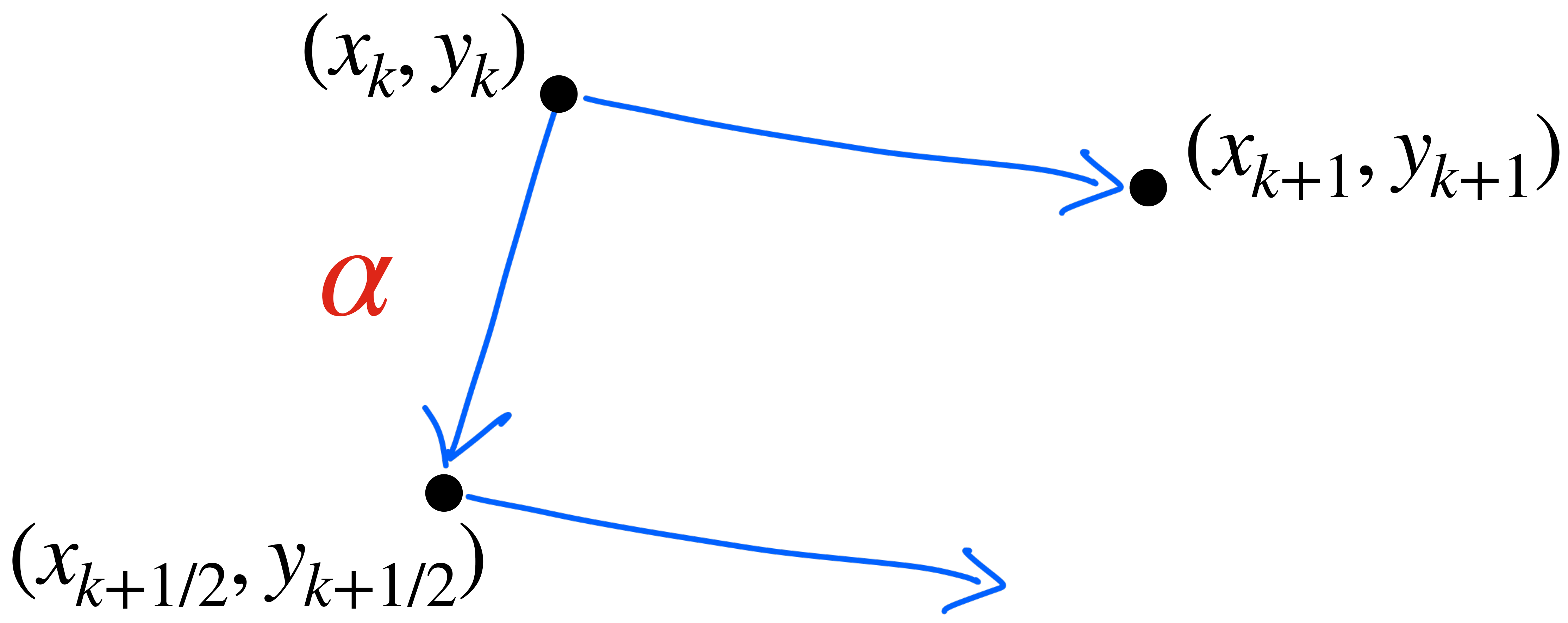
↳ <u>Mirror-prox:</u> rough solution = a gradient step , $\alpha = L$



Poster #212

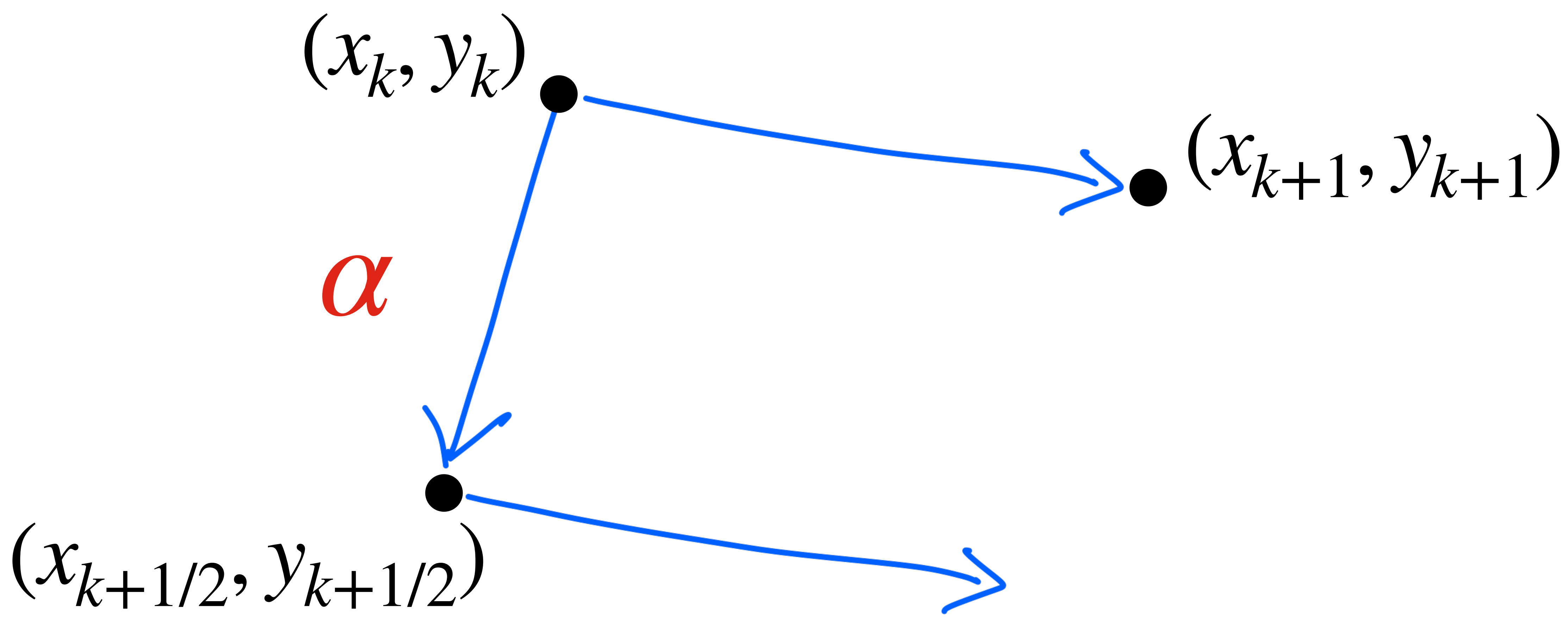
Variance reduction framework

Method	# of iterations	cost per iteration
<u>Nemirovski's "conceptual prox-method"</u> $(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to $f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2$ $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox $+ \, T_{\text{exact}}$
↳ <u>Mirror-prox:</u> rough solution = a gradient step , $\alpha = L$	$\frac{\alpha}{\epsilon} = \frac{L}{\epsilon}$	



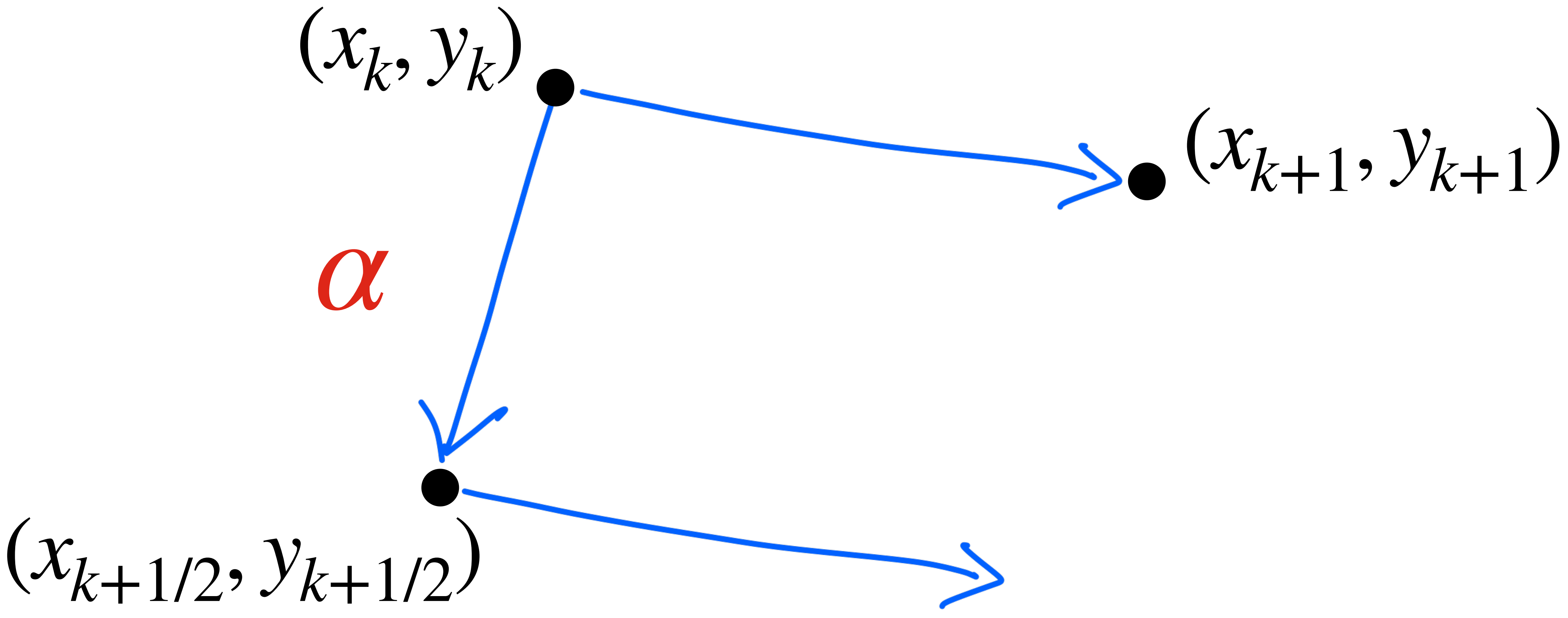
Variance reduction framework

Method	# of iterations	cost per iteration
<u>Nemirovski's "conceptual prox-method"</u> $(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to $f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2$ $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox $+ T_{\text{exact}}$
↳ <u>Mirror-prox:</u> rough solution = a gradient step , $\alpha = L$	$\frac{\alpha}{\epsilon} = \frac{L}{\epsilon}$	T_{exact}



Variance reduction framework

Method	# of iterations	cost per iteration
<u>Nemirovski's "conceptual prox-method"</u> $(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to $f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2$ $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox $+ \, T_{\text{exact}}$
↳ <u>Mirror-prox:</u> rough solution = a gradient step, $\alpha = L$	$\frac{\alpha}{\epsilon} = \frac{L}{\epsilon}$	T_{exact}



Total runtime

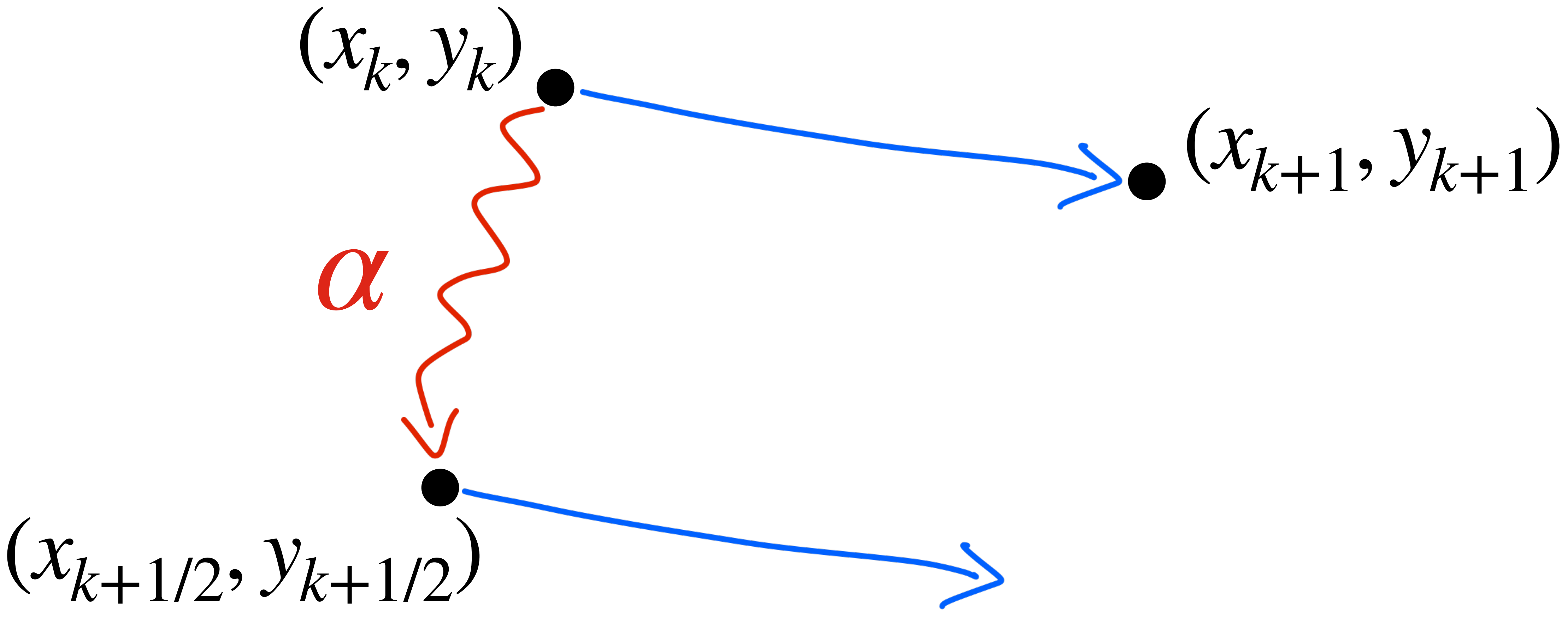
$$\frac{L}{\epsilon} \cdot \underset{(=n^2)}{T_{\text{exact}}}$$

Variance reduction framework

Method	# of iterations	cost per iteration
<u>Nemirovski's "conceptual prox-method"</u> $(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to $f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2$ $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox $+ T_{\text{exact}}$
<u>Mirror-prox:</u> rough solution = a gradient step , $\alpha = L$	$\frac{\alpha}{\epsilon} = \frac{L}{\epsilon}$	T_{exact}

<u>Our appoach:</u>

rough solution = **centered stochastic gradient steps**



Total runtime

$$\frac{L}{\epsilon}\sqrt{T_{\text{exact}} T_{\text{stoch}}}$$

$(= n^{3/2})$

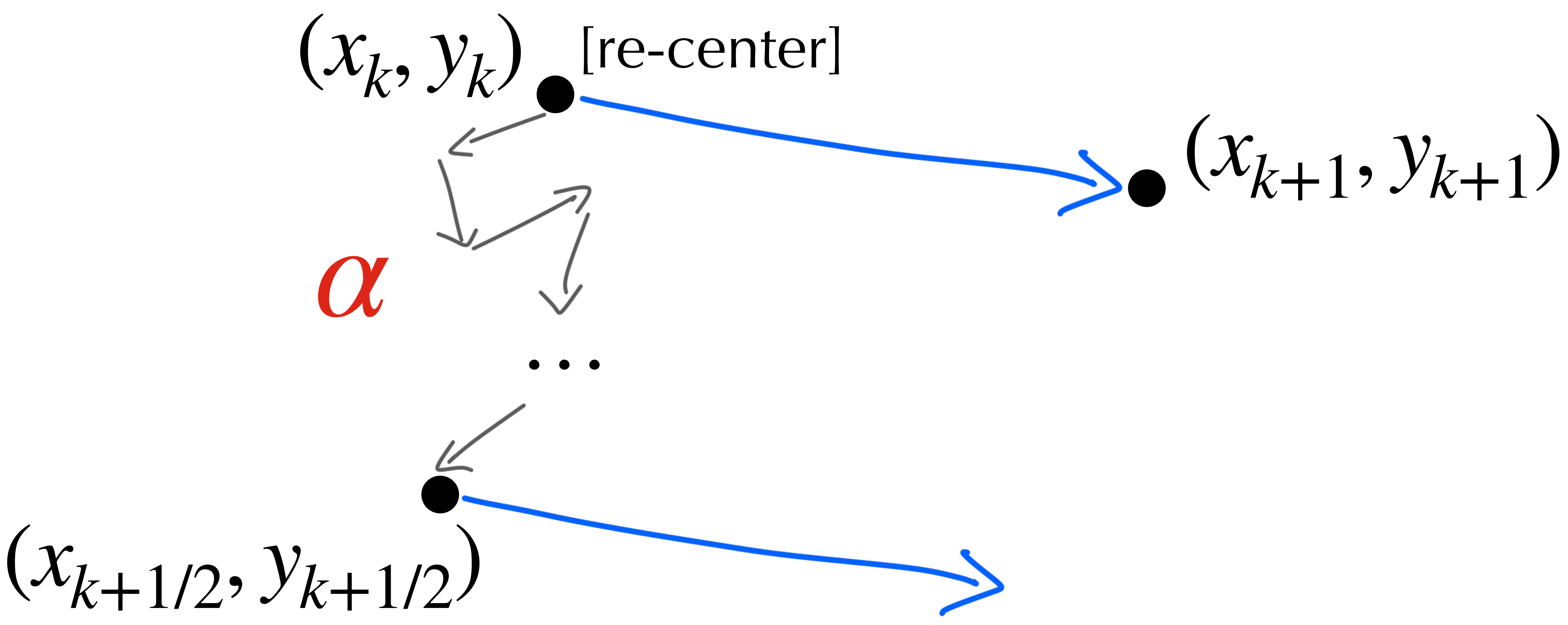
Poster #212

Variance reduction framework

Method	# of iterations	cost per iteration
<u>Nemirovski's "conceptual prox-method"</u> $(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to $f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2$ $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox $+ T_{\text{exact}}$
<u>Mirror-prox:</u> rough solution = a gradient step , $\alpha = L$	$\frac{\alpha}{\epsilon} = \frac{L}{\epsilon}$	T_{exact}

<u>Our appoach:</u>

rough solution = **centered stochastic gradient steps**



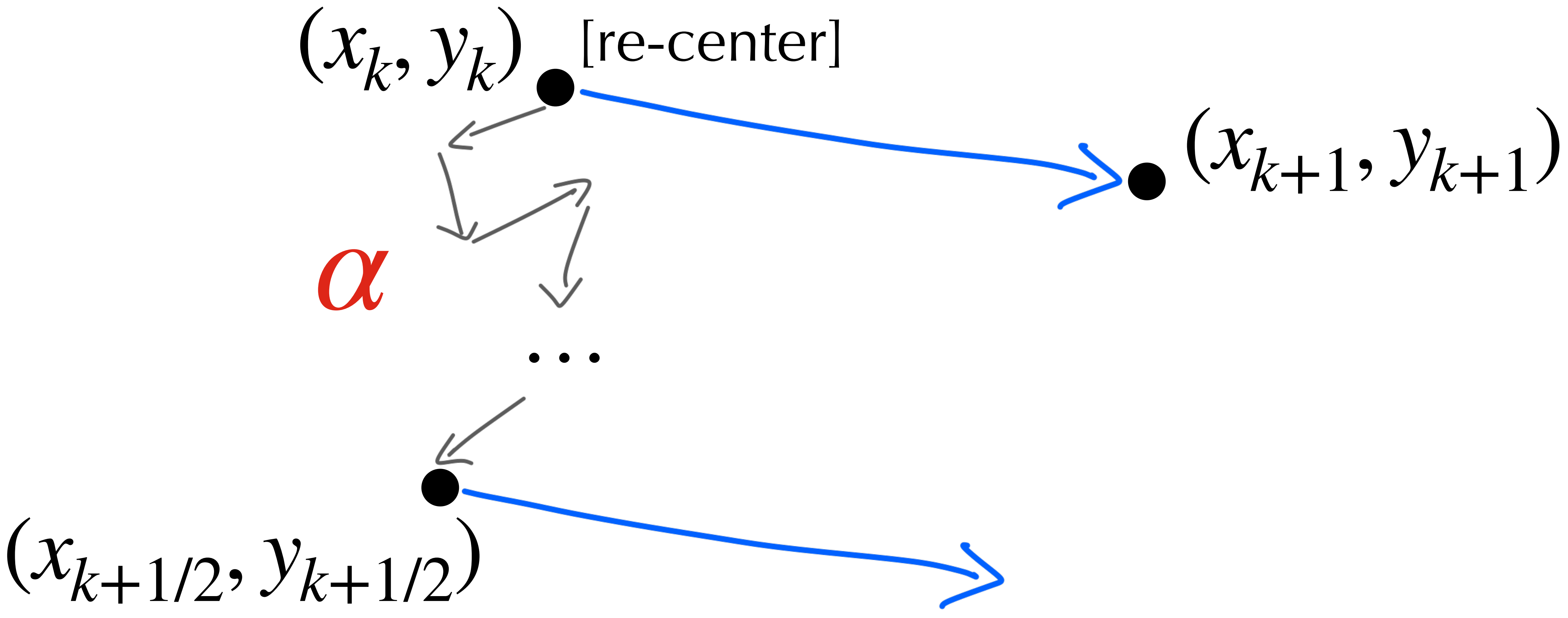
Total runtime

$$\frac{L}{\epsilon}\sqrt{T_{\text{exact}} T_{\text{stoch}}}$$

$(= n^{3/2})$

Variance reduction framework

Method	# of iterations	cost per iteration
Nemirovski's "conceptual prox-method" $(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to $f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2$ $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox $+ T_{\text{exact}}$
Mirror-prox: rough solution = a gradient step , $\alpha = L$	$\frac{\alpha}{\epsilon} = \frac{L}{\epsilon}$	T_{exact}
Our appoach: rough solution = **centered stochastic gradient steps**		$T_{\text{stoch}} \cdot \frac{L^2}{\alpha^2} + T_{\text{exact}}$ (main technical development)



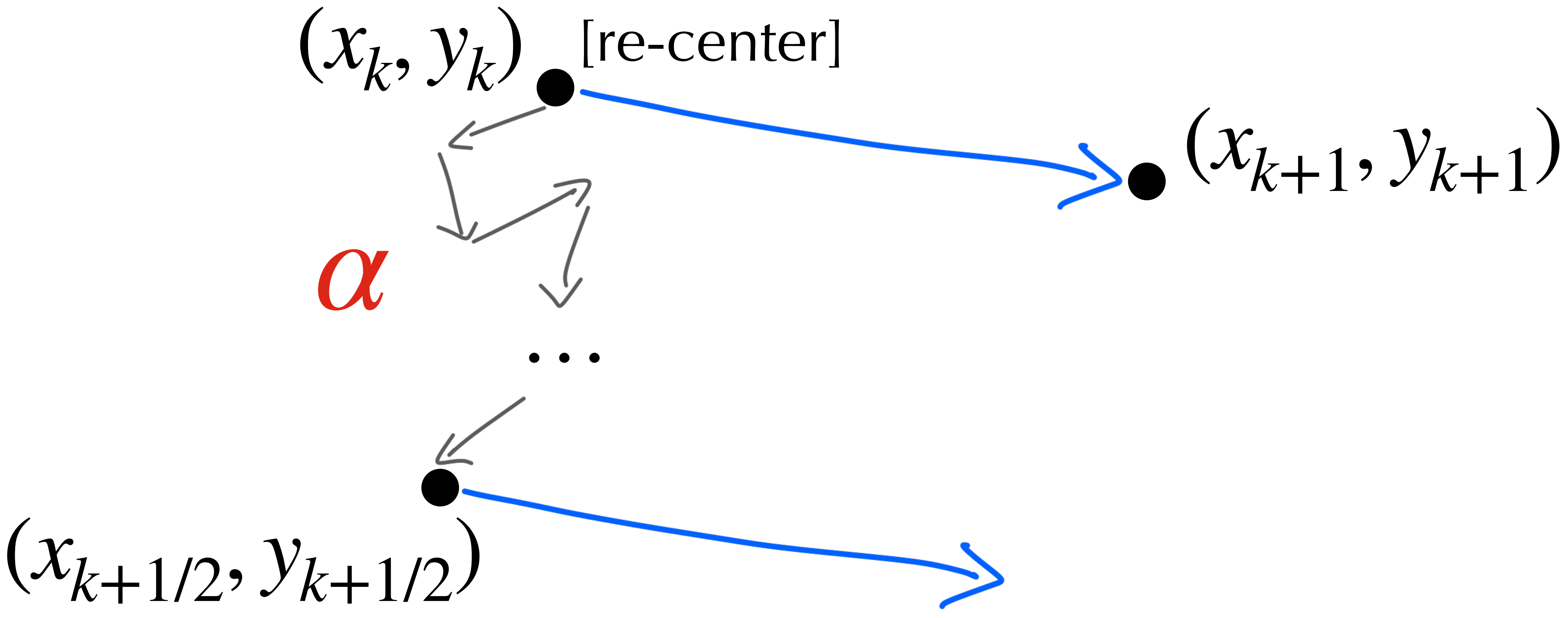
Total runtime

$$\frac{L}{\epsilon}\sqrt{T_{\text{exact}} T_{\text{stoch}}}$$

$(= n^{3/2})$

Variance reduction framework

Method	# of iterations	cost per iteration
<u>Nemirovski's "conceptual prox-method"</u> $(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to $f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2$ $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox $+\ T_{\text{exact}}$
<u>Mirror-prox:</u> rough solution = a gradient step , $\alpha = L$	$\frac{\alpha}{\epsilon} = \frac{L}{\epsilon}$	T_{exact}
<u>Our appoach:</u> rough solution = **centered stochastic gradient steps** , $\alpha = L\sqrt{\frac{T_{\text{stoch}}}{T_{\text{exact}}}}$		$T_{\text{stoch}} \cdot \frac{L^2}{\alpha^2} + T_{\text{exact}}$ (main technical development)



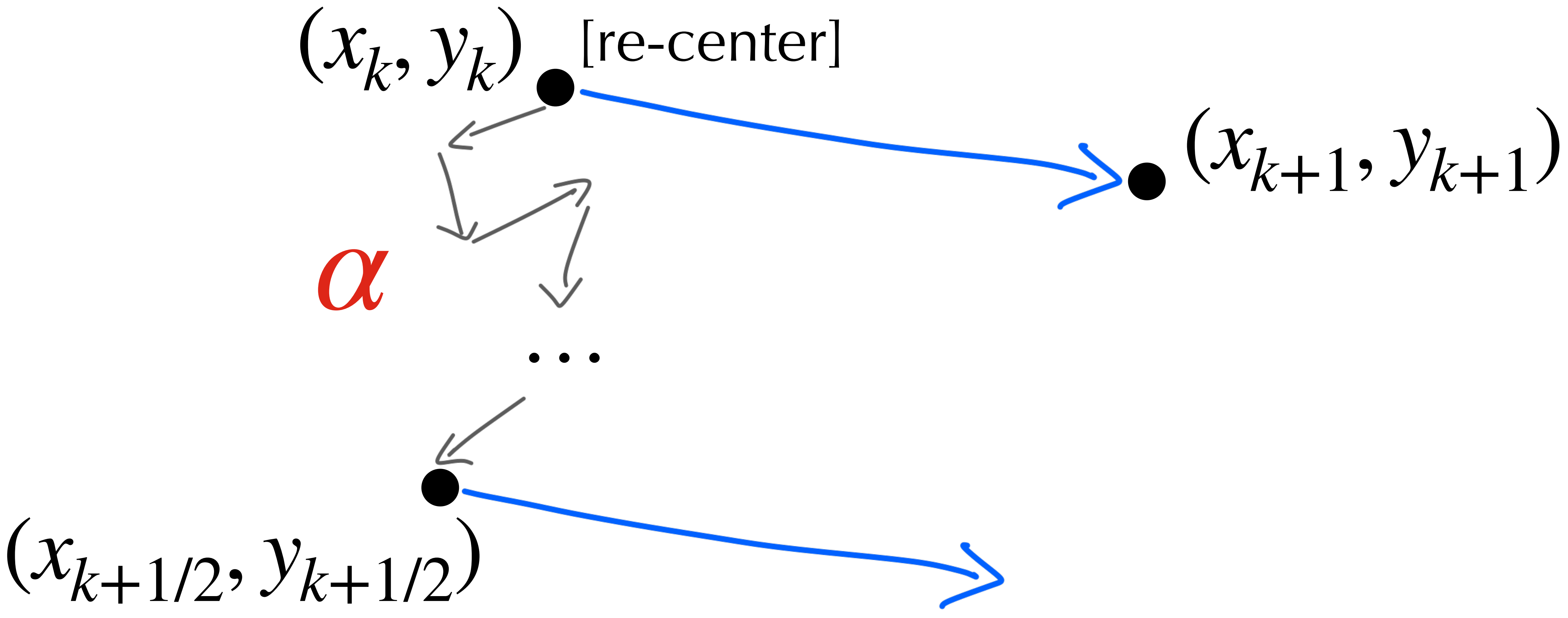
Total runtime

$$\frac{L}{\epsilon}\sqrt{T_{\text{exact}} T_{\text{stoch}}}$$

$(= n^{3/2})$

Variance reduction framework

Method	# of iterations	cost per iteration
Nemirovski's "conceptual prox-method" $(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to $f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2$ $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox $+ T_{\text{exact}}$
Mirror-prox: rough solution = a gradient step , $\alpha = L$	$\frac{\alpha}{\epsilon} = \frac{L}{\epsilon}$	T_{exact}
Our appoach: rough solution = **centered stochastic gradient steps** , $\alpha = L\sqrt{\frac{T_{\text{stoch}}}{T_{\text{exact}}}}$		$T_{\text{stoch}} \cdot \frac{L^2}{\alpha^2} + T_{\text{exact}}$ T_{exact} (main technical development)



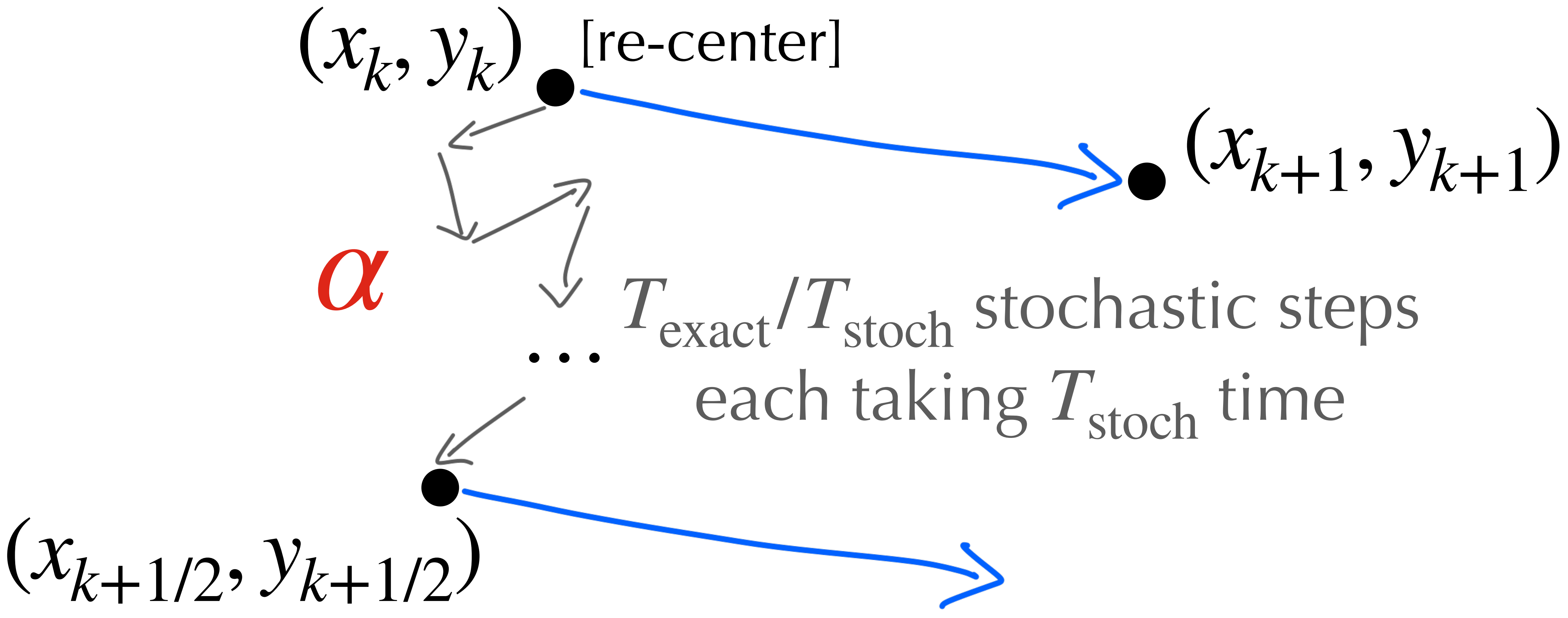
Total runtime

$$\frac{L}{\epsilon}\sqrt{T_{\text{exact}} T_{\text{stoch}}}$$

$(= n^{3/2})$

Variance reduction framework

Method	# of iterations	cost per iteration
<u>Nemirovski's "conceptual prox-method"</u> $(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to $f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2$ $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox $+ \, T_{\text{exact}}$
<u>Mirror-prox:</u> rough solution = a gradient step , $\alpha = L$	$\frac{\alpha}{\epsilon} = \frac{L}{\epsilon}$	T_{exact}
<u>Our appoach:</u> rough solution = **centered stochastic gradient steps** , $\alpha = L\sqrt{\frac{T_{\text{stoch}}}{T_{\text{exact}}}}$		T_{exact} $T_{\text{stoch}} \cdot \frac{L^2}{\alpha^2} + T_{\text{exact}}$ (main technical development)



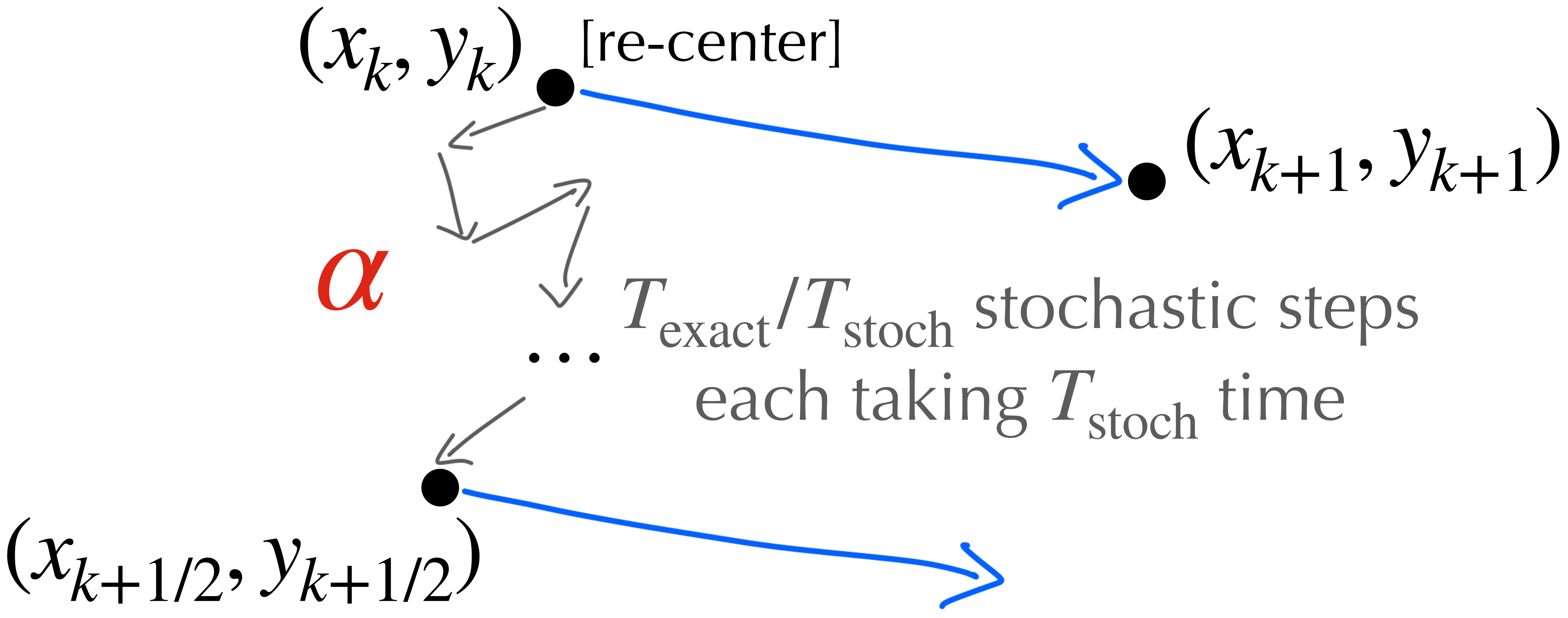
Total runtime

$$\frac{L}{\epsilon}\sqrt{T_{\text{exact}} T_{\text{stoch}}}$$

$(= n^{3/2})$

Variance reduction framework

Method	# of iterations	cost per iteration
Nemirovski's "conceptual prox-method" $(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to $f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2$ $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox $+ T_{\text{exact}}$
Mirror-prox: rough solution = a gradient step , $\alpha = L$	$\frac{\alpha}{\epsilon} = \frac{L}{\epsilon}$	T_{exact}
Our appoach: rough solution = **centered stochastic gradient steps** , $\alpha = L\sqrt{\frac{T_{\text{stoch}}}{T_{\text{exact}}}}$	$\frac{\alpha}{\epsilon} = \frac{L}{\epsilon}\sqrt{\frac{T_{\text{stoch}}}{T_{\text{exact}}}}$	$T_{\text{stoch}} \cdot \frac{L^2}{\alpha^2}$ T_{exact} $+ T_{\text{exact}}$ (main technical development)



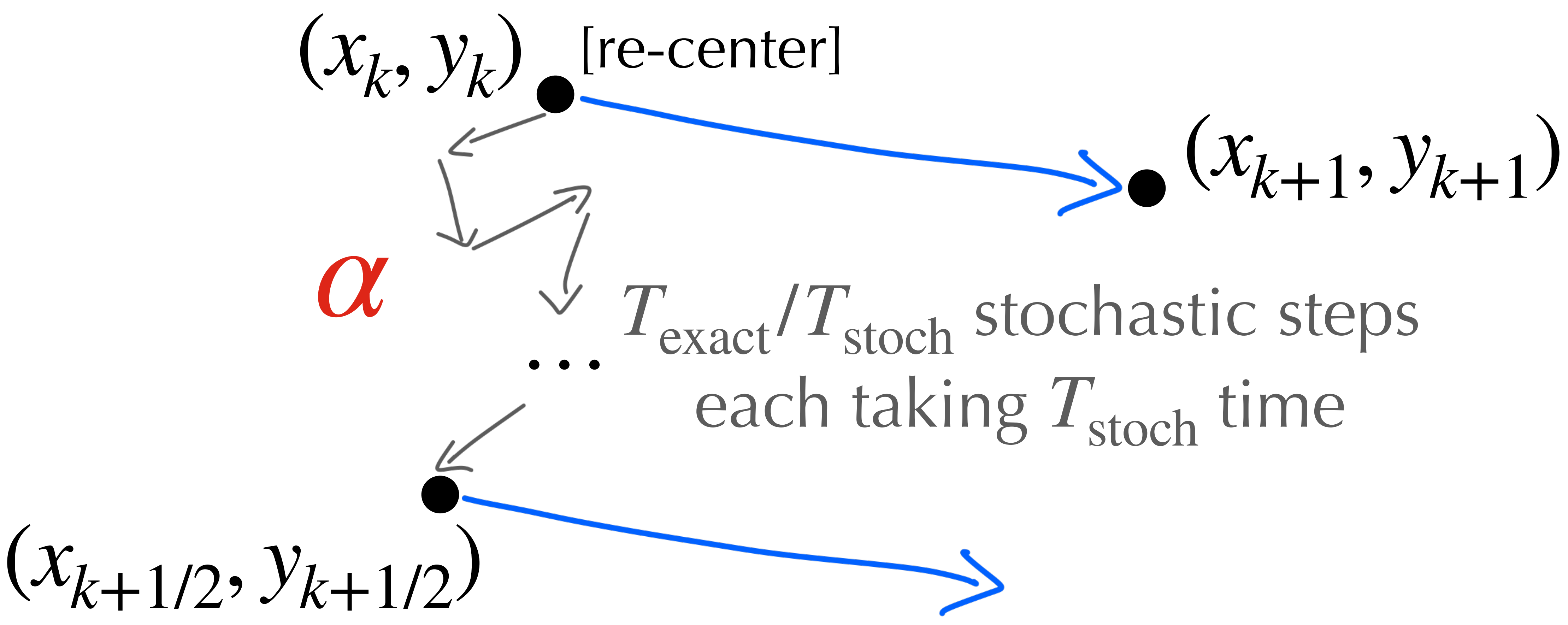
Total runtime

$$\frac{L}{\epsilon}\sqrt{T_{\text{exact}} T_{\text{stoch}}}$$

$(= n^{3/2})$

Variance reduction framework

Method	# of iterations	cost per iteration
<u>Nemirovski's "conceptual prox-method"</u> $(x_{k+1/2}, y_{k+1/2}) \leftarrow$ rough solution to $f(x,y) + \frac{\alpha}{2}\|x - x_k\|^2 - \frac{\alpha}{2}\|y - y_k\|^2$ $(x_{k+1}, y_{k+1}) \leftarrow$ extra-gradient step (exact gradient)	$\frac{\alpha}{\epsilon}$	cost of rough prox $+ T_{\text{exact}}$
<u>Mirror-prox:</u> rough solution = a gradient step , $\alpha = L$	$\frac{\alpha}{\epsilon} = \frac{L}{\epsilon}$	T_{exact}
<u>Our appoach:</u> rough solution = **centered stochastic gradient steps** , $\alpha = L\sqrt{\frac{T_{\text{stoch}}}{T_{\text{exact}}}}$	$\frac{\alpha}{\epsilon} = \frac{L}{\epsilon}\sqrt{\frac{T_{\text{stoch}}}{T_{\text{exact}}}}$	$T_{\text{stoch}} \cdot \frac{L^2}{\alpha^2}$ T_{exact} $+ T_{\text{exact}}$ (main technical development)



Total runtime

$$\frac{L}{\epsilon}\sqrt{T_{\text{exact}} T_{\text{stoch}}}$$

$(= n^{3/2})$

Summary

Poster #212

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$$

Summary

Poster #212

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$$

Centered gradient estimator

$$\mathrm{Var}\, g_{z_0}(z) \le L^2 \|z - z_0\|^2$$

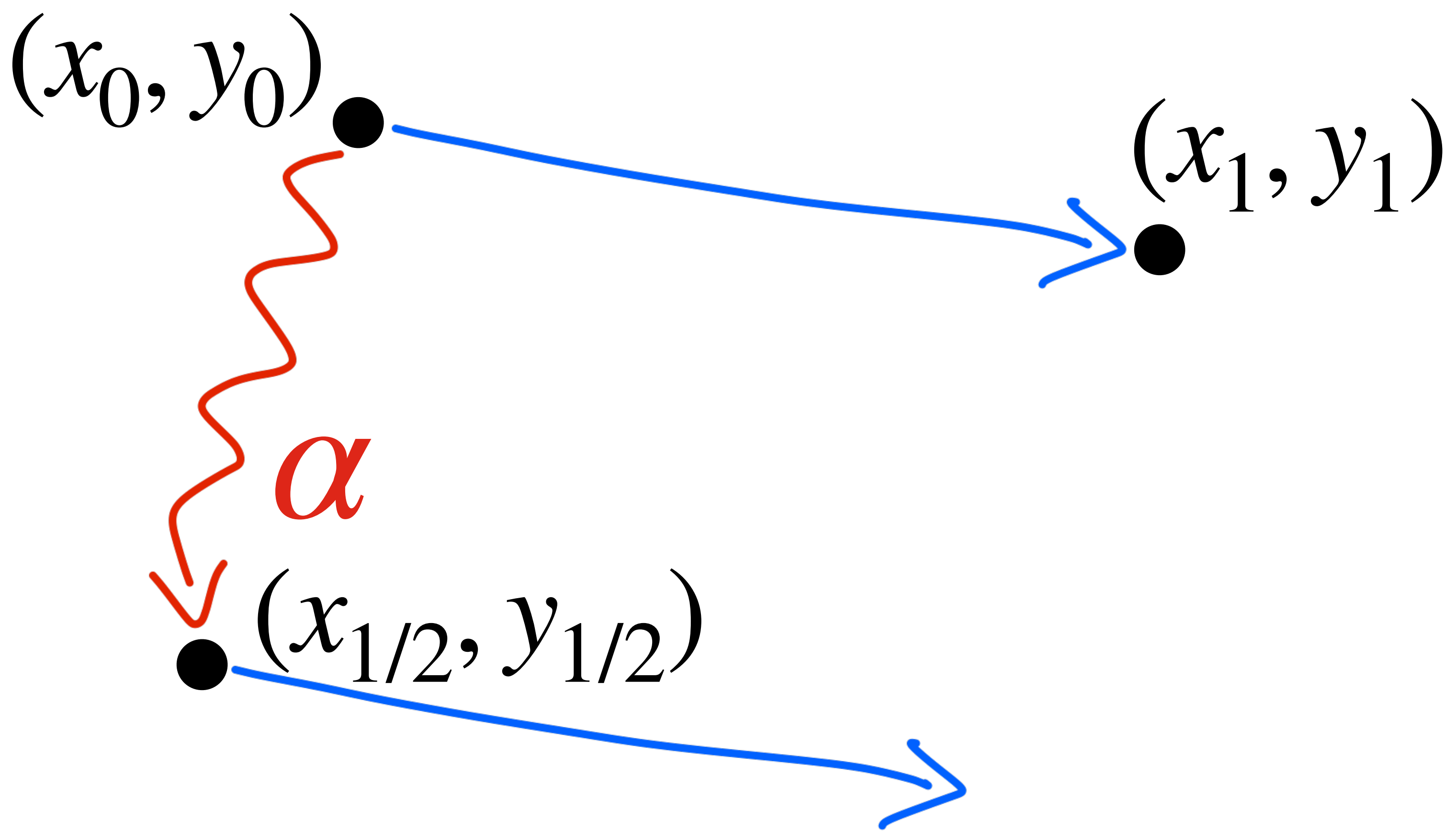
Summary

Poster #212

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$$

Centered gradient estimator

$$\mathrm{Var}\, g_{z_0}(z) \le L^2 \|z - z_0\|^2$$



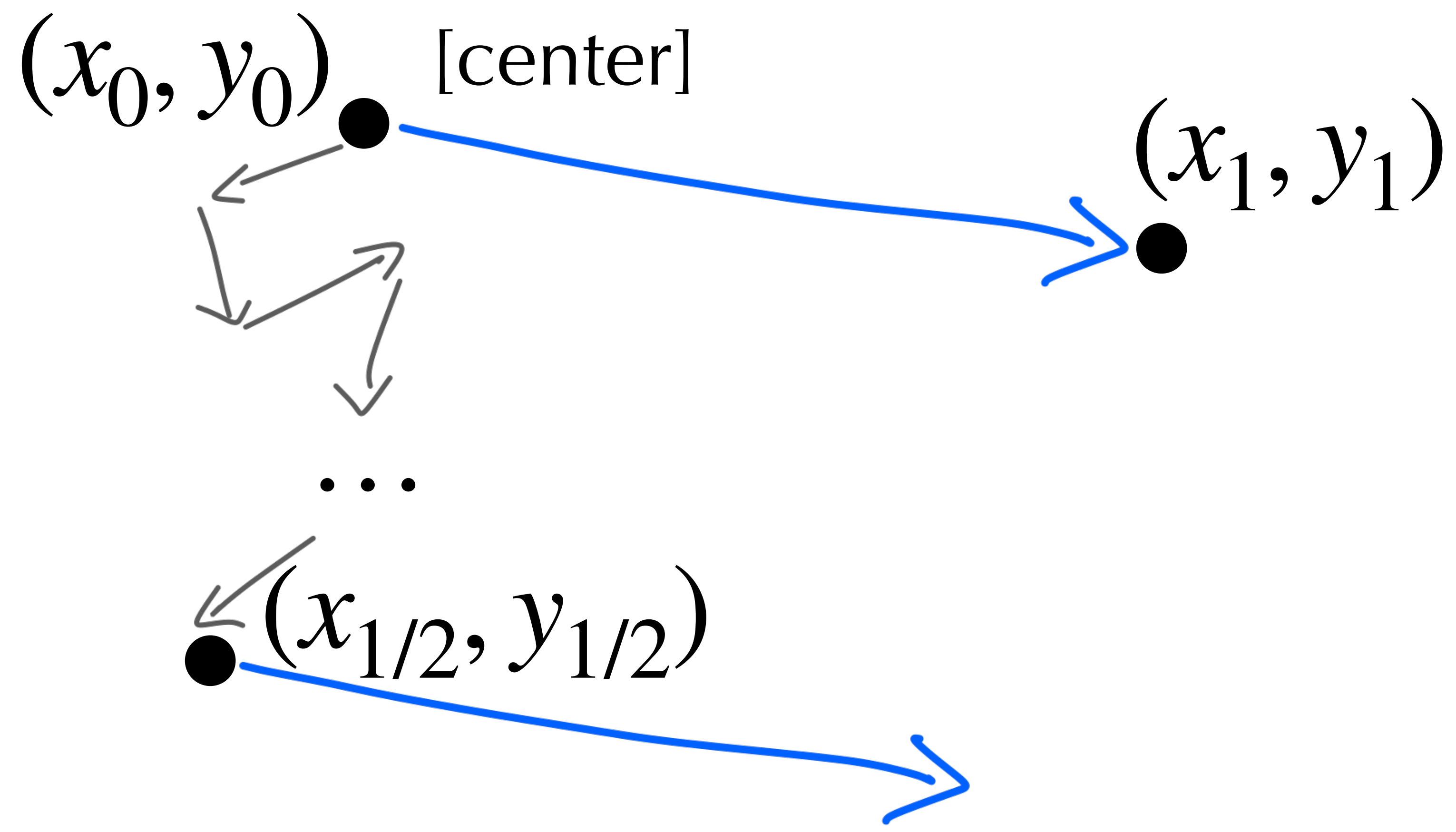
Summary

Poster #212

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$$

Centered gradient estimator

$$\mathrm{Var}\, g_{z_0}(z) \le L^2 \|z - z_0\|^2$$



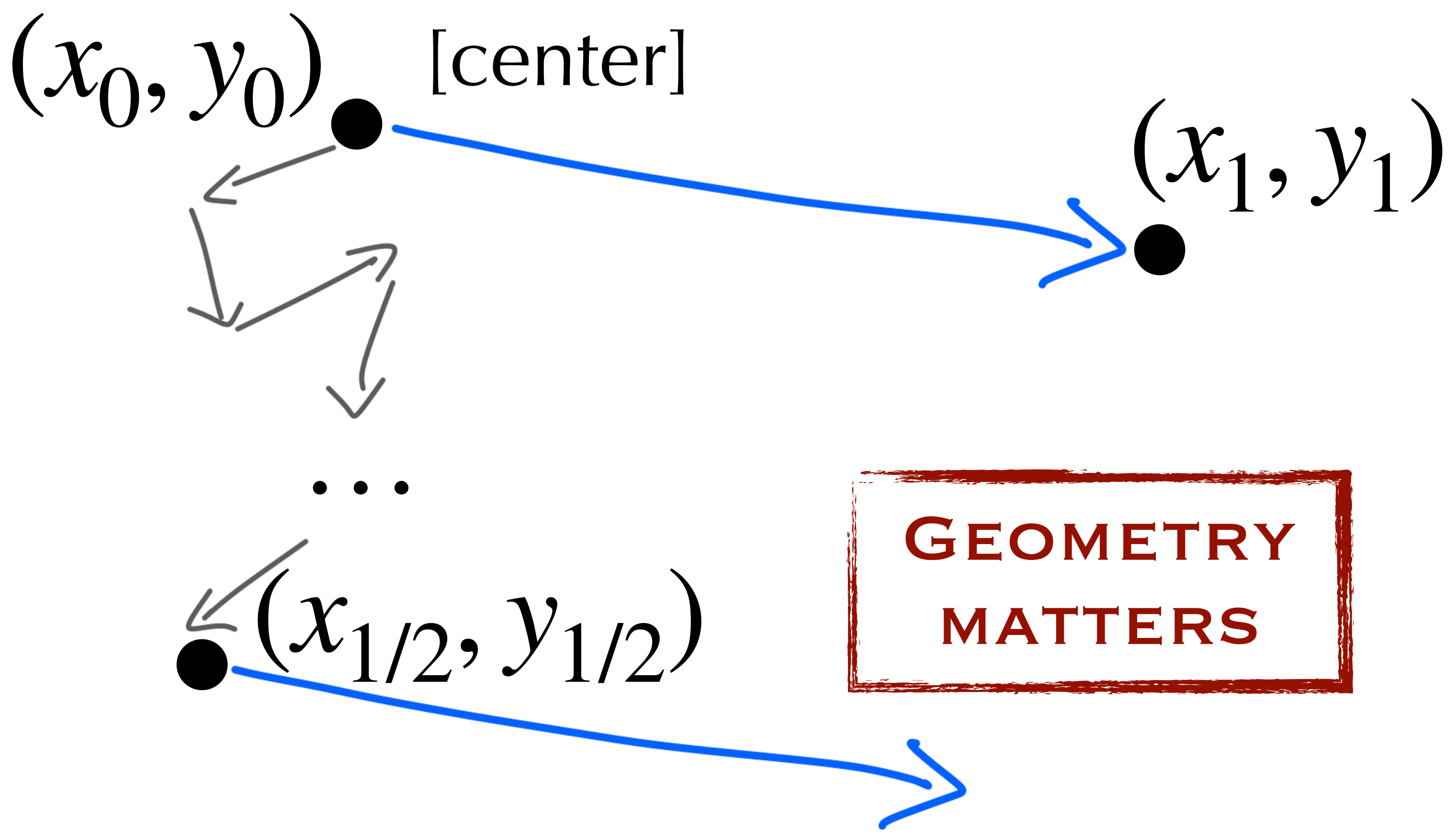
Summary

Poster #212

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$$

Centered gradient estimator

$$\operatorname{Var} g_{z_0}(z) \leq L^2 \|z - z_0\|^2$$



GEOMETRY MATTERS

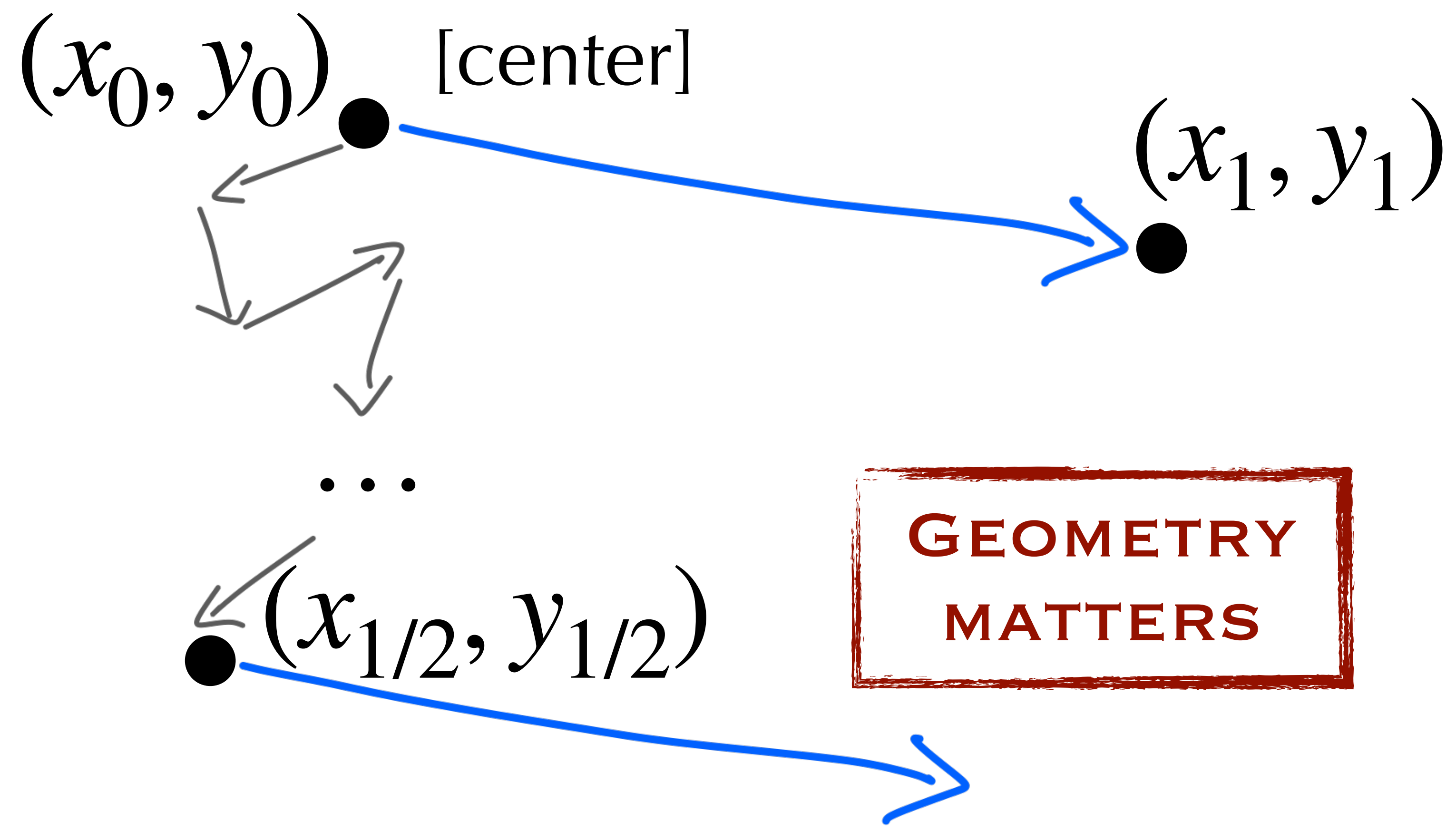
Summary

Poster #212

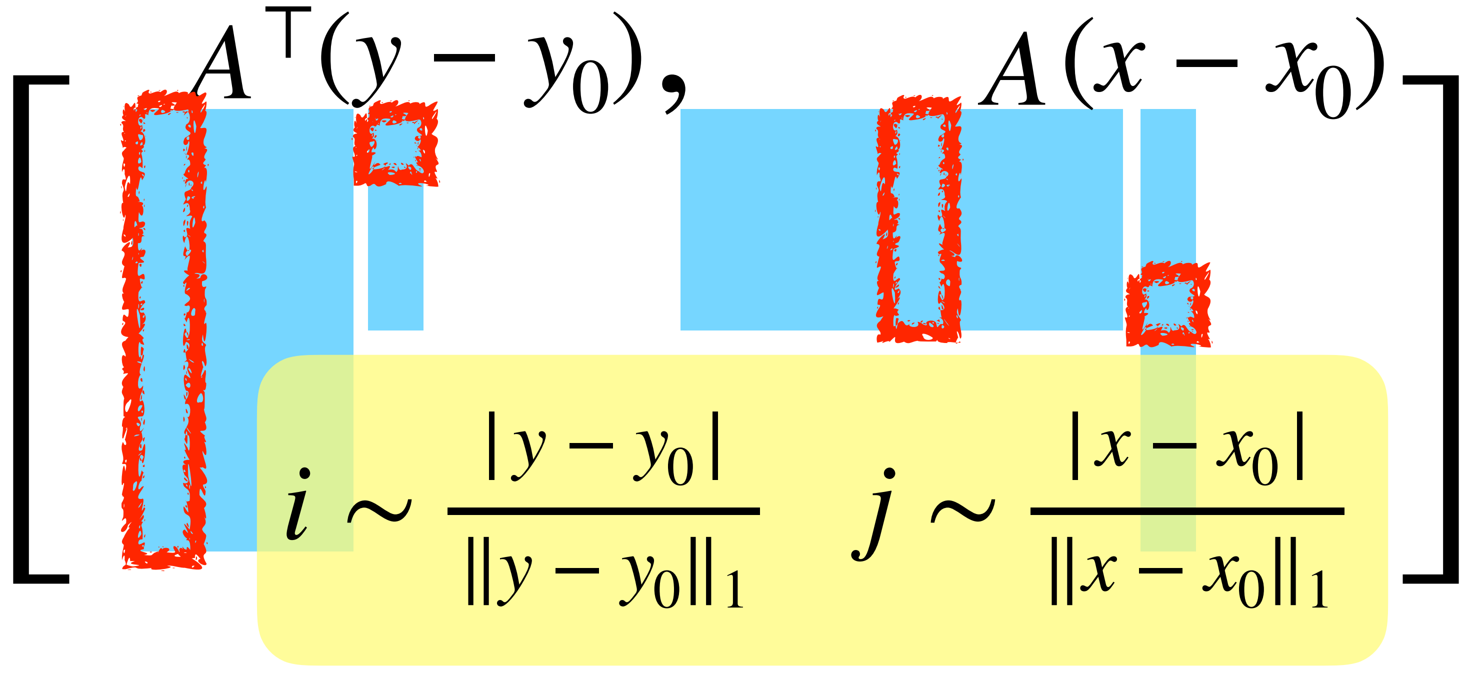
$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x,y)$$

Centered gradient estimator

$$\mathrm{Var}\, g_{z_0}(z) \le L^2\|z-z_0\|^2$$



sampling from the difference



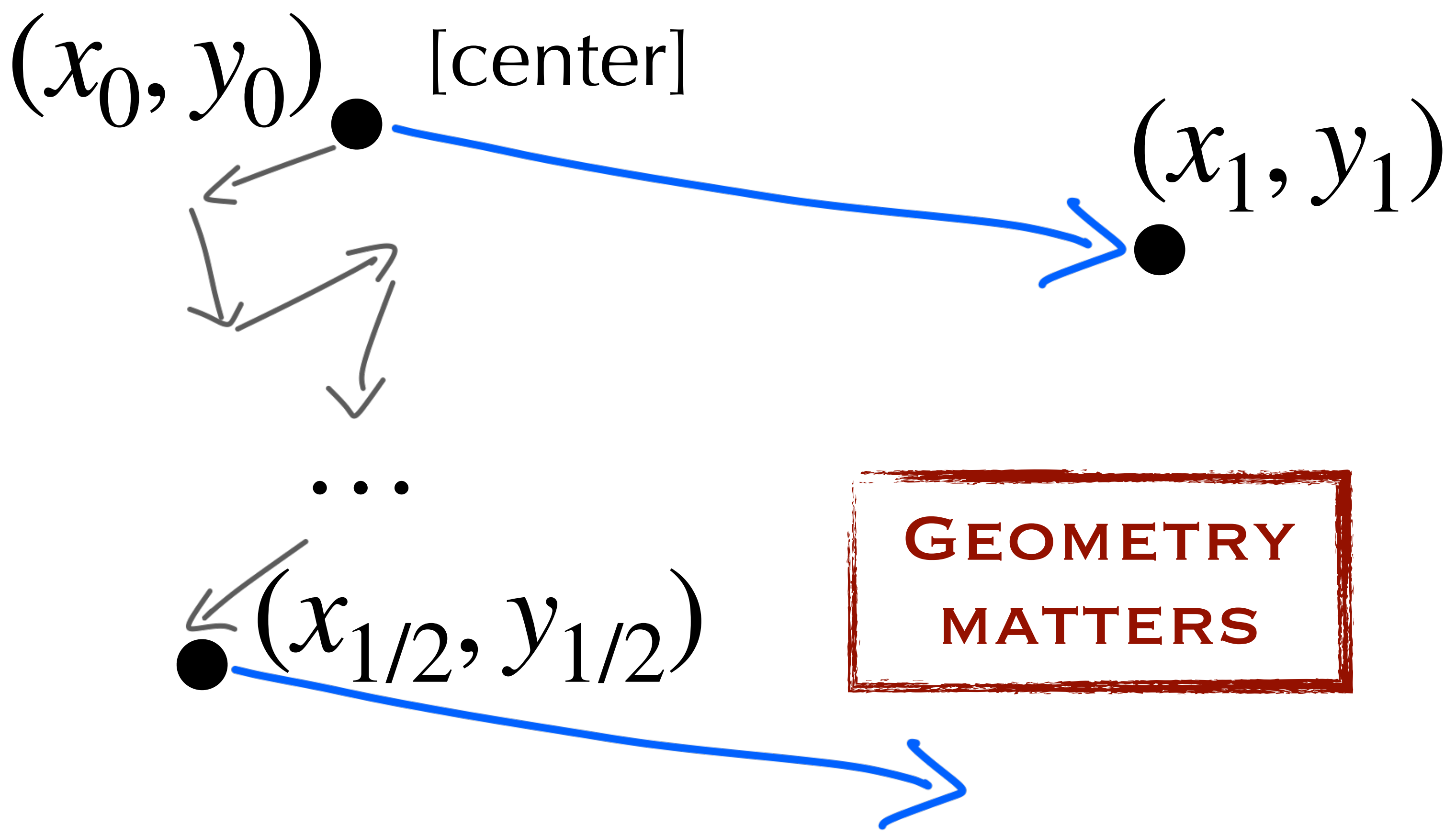
Poster #212

Summary

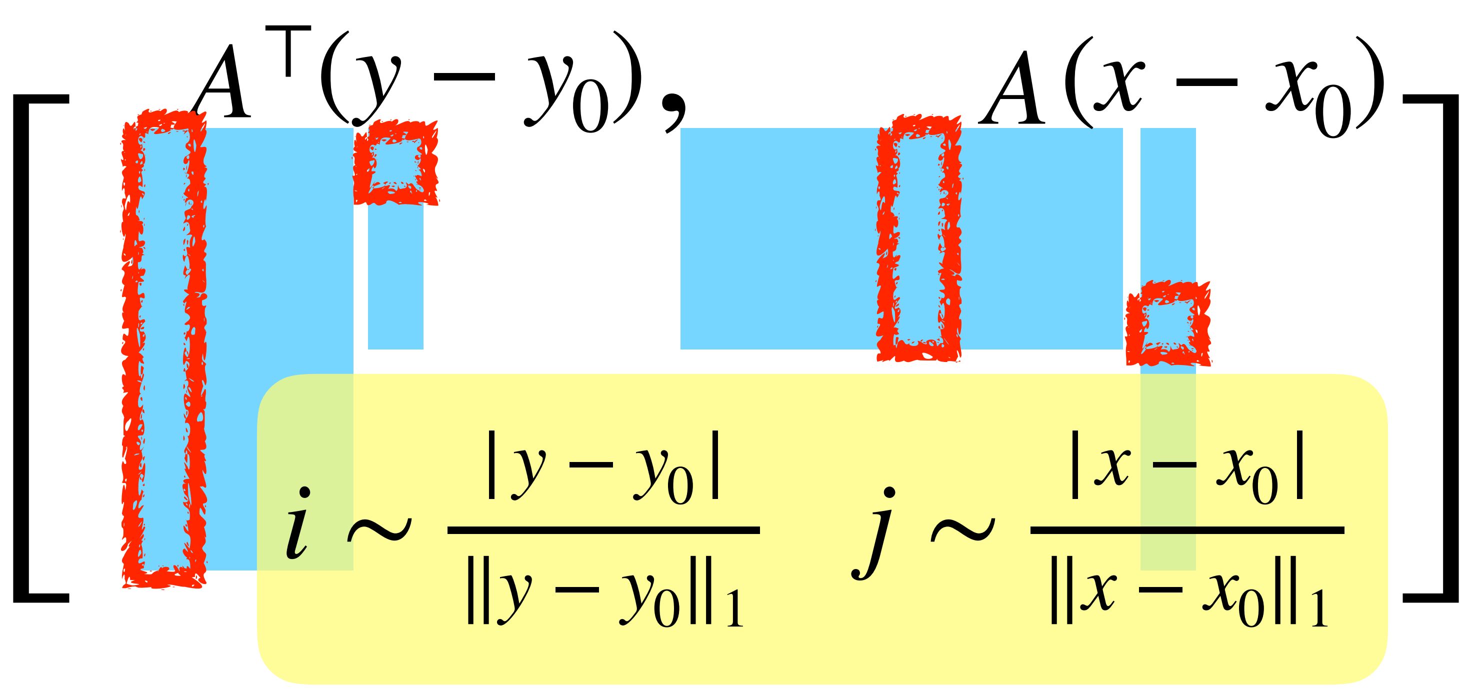
$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$$

Centered gradient estimator

$$\mathrm{Var}\, g_{z_0}(z) \le L^2 \|z - z_0\|^2$$



sampling from the difference



Exact gradient

(Nemirovski `04, Nesterov `07)

$$n^2 \cdot \frac{L}{\epsilon}$$

Variance reduction

(our approach)

$$n^2 + n^{3/2} \cdot \frac{L}{\epsilon}$$

Stochastic gradient

(GK95, CHW10)

$$n \cdot \frac{L^2}{\epsilon^2}$$

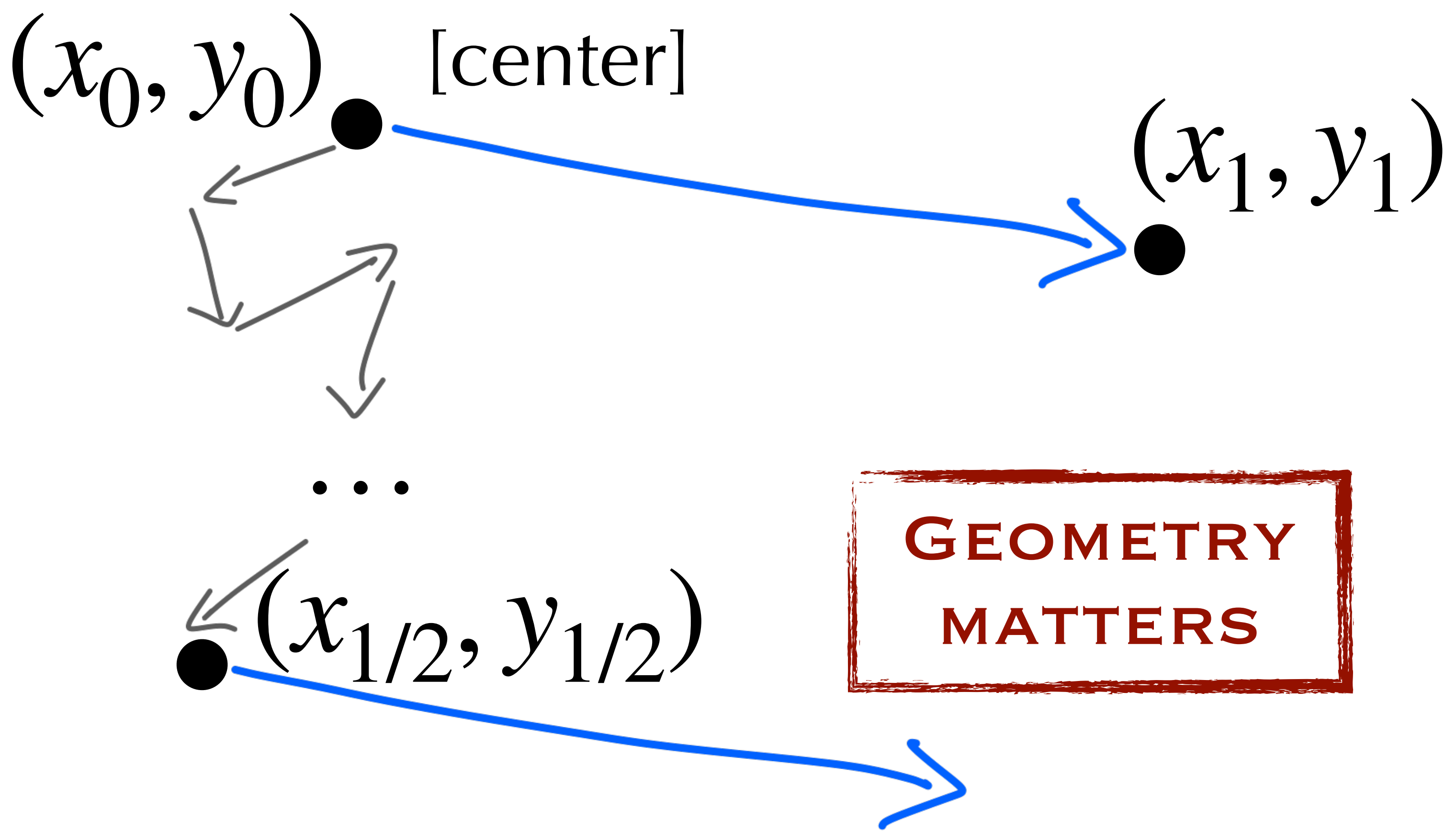
Summary

Poster #212

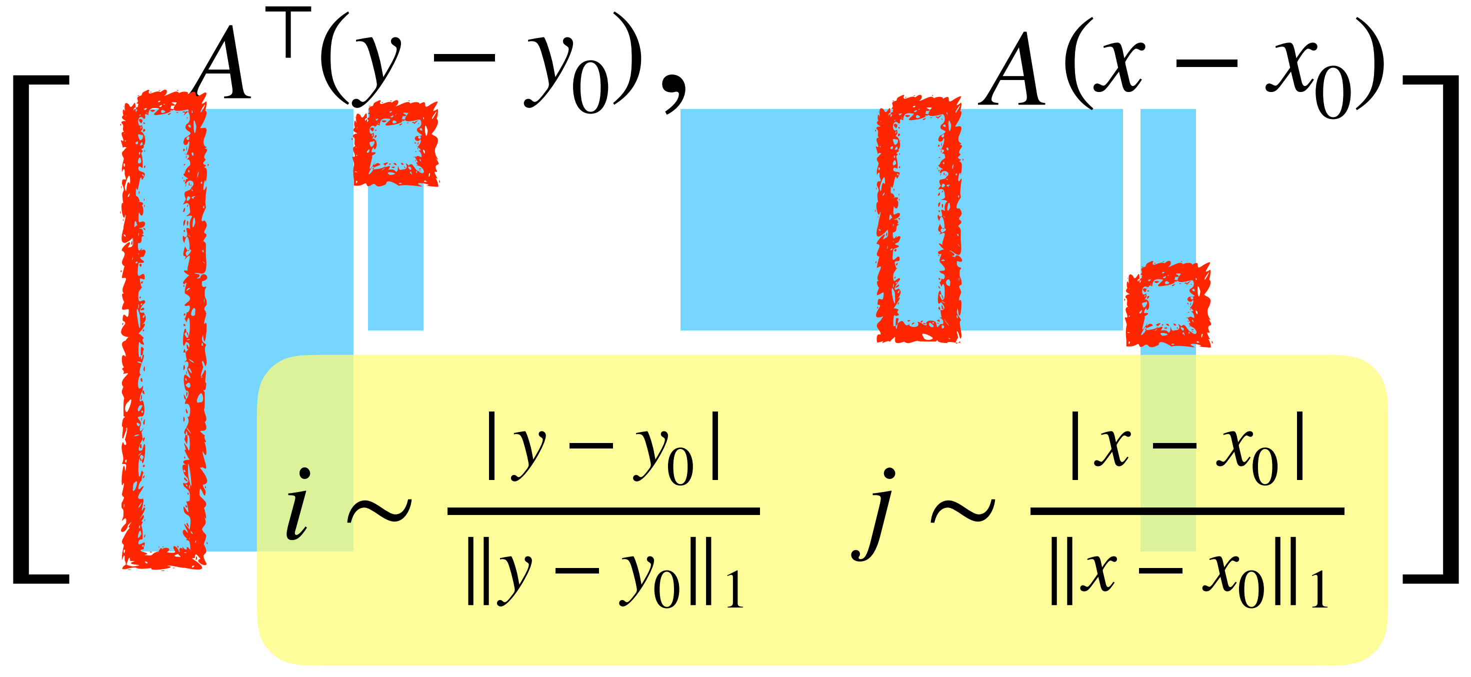
$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$$

Centered gradient estimator

$$\mathrm{Var}\, g_{z_0}(z) \le L^2 \|z - z_0\|^2$$



sampling from the difference



Exact gradient

(Nemirovski \`04, Nesterov \`07)

$$n^2 \cdot \frac{L}{\epsilon}$$

Variance reduction

(our approach)

$$n^2 + n^{3/2} \cdot \frac{L}{\epsilon}$$

Stochastic gradient

(GK95, CHW10)

$$n \cdot \frac{L^2}{\epsilon^2}$$

Image credit: Chawit Waewsawangwong

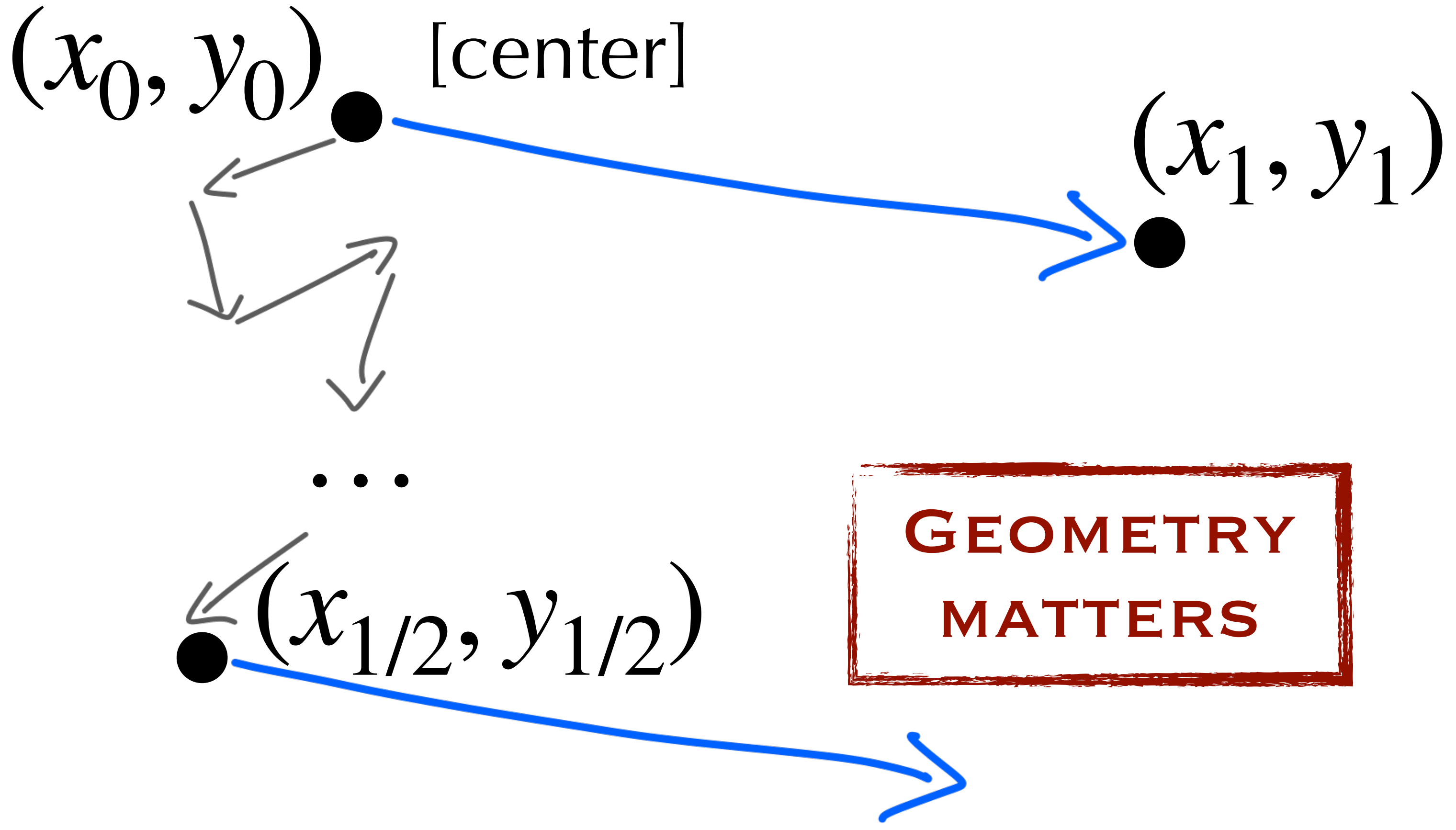
Summary

Poster #212

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y)$$

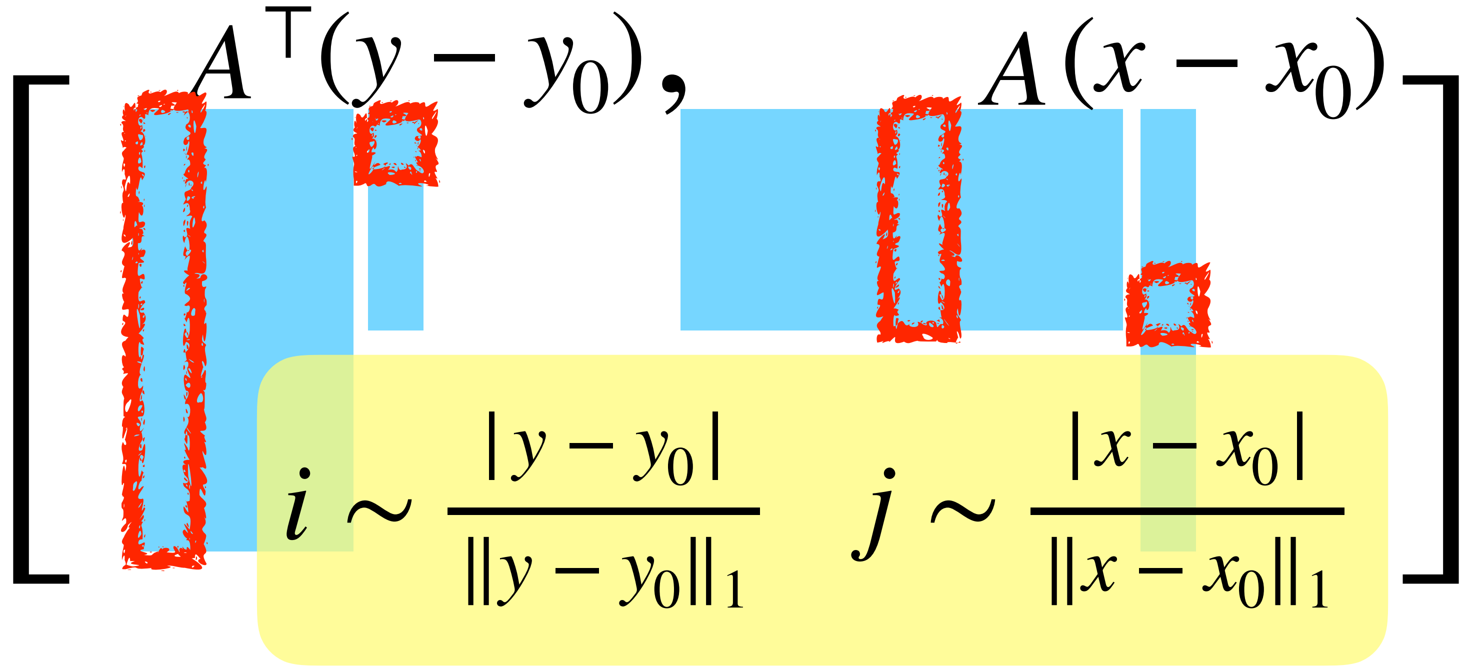
Centered gradient estimator

$$\mathrm{Var}\, g_{z_0}(z) \le L^2 \|z - z_0\|^2$$



GEOMETRY MATTERS

sampling from the difference



Exact gradient

(Nemirovski `04, Nesterov `07)

$$n^2 \cdot \frac{L}{\epsilon}$$

Variance reduction

(our approach)

$$n^2 + n^{3/2} \cdot \frac{L}{\epsilon}$$

Stochastic gradient

(GK95, CHW10)

$$n \cdot \frac{L^2}{\epsilon^2}$$

Image credit: Chawit Waewsawangwong

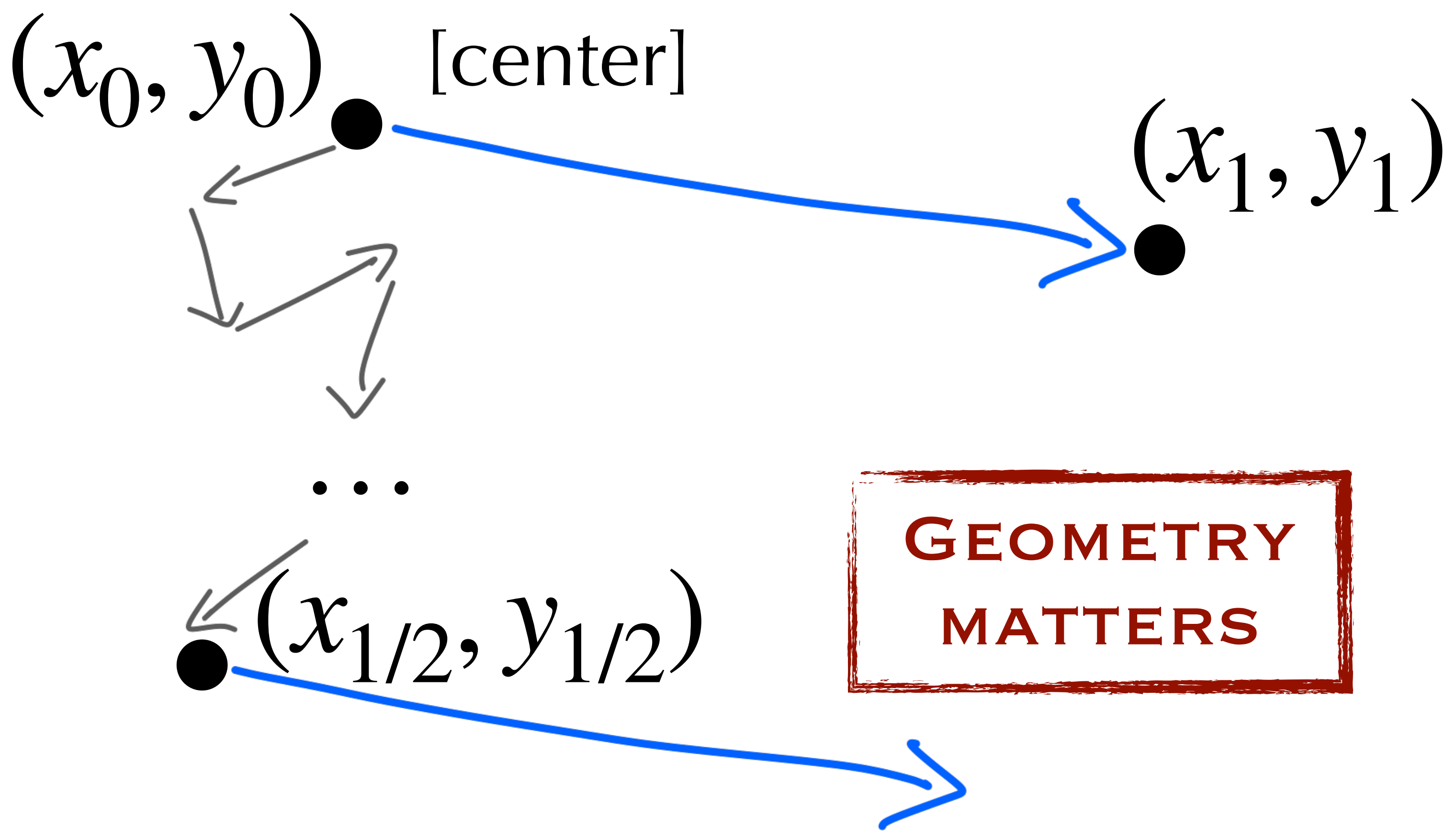
Summary

Poster #212

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$$

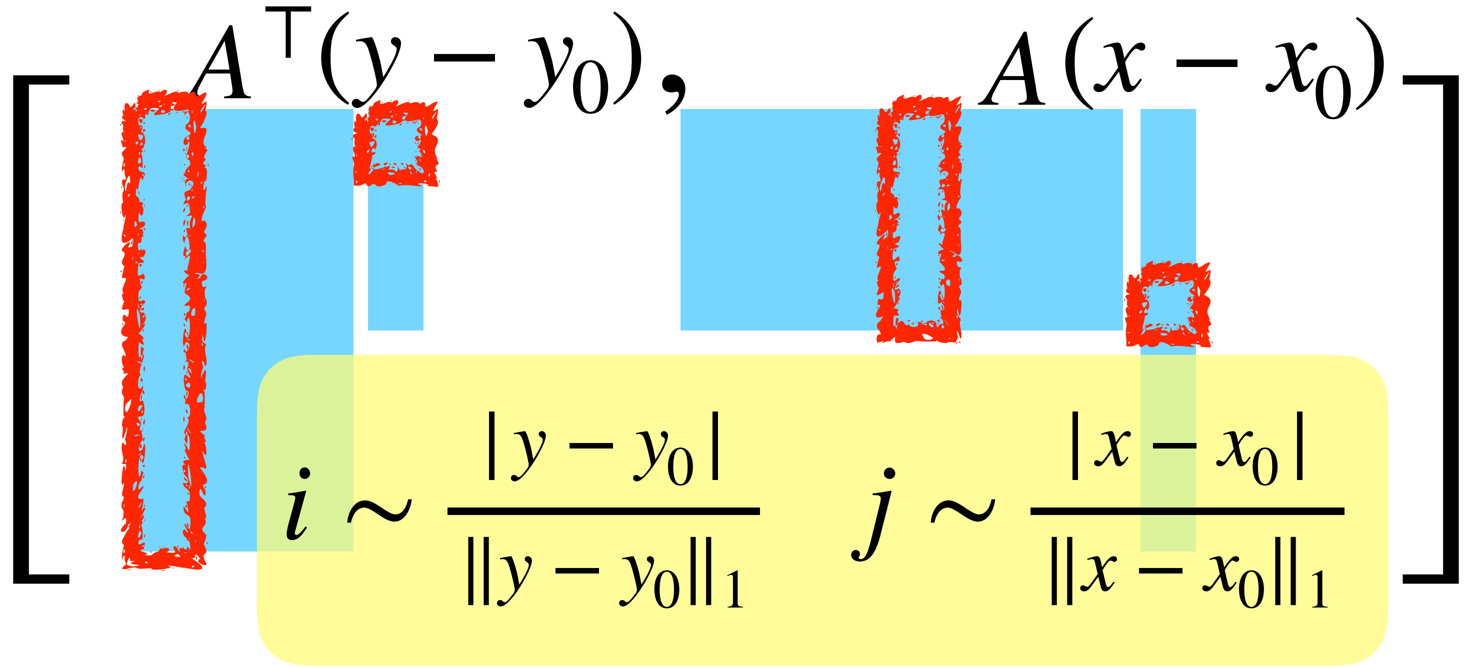
Centered gradient estimator

$$\mathrm{Var}\, g_{z_0}(z) \le L^2 \|z - z_0\|^2$$



GEOMETRY MATTERS

sampling from the difference



Exact gradient

(Nemirovski \`04, Nesterov \`07)

$$n^2 \cdot \frac{L}{\epsilon}$$

Variance reduction

(our approach)

$$n^2 + n^{3/2} \cdot \frac{L}{\epsilon}$$

Stochastic gradient

(GK95, CHW10)

$$n \cdot \frac{L^2}{\epsilon^2}$$

VR always better

VR better for $\Omega(1)$ passes over data

GEOMETRY MATTERS

Image credit: Chawit Waewsawangwong