Advanced Topic: Dual Averaging

Kevin Waugh
waugh@cs.cmu.edu

Yet another gradient method (YAGM)

Outline

- Review: Subgradient Method
- Derive Dual Averaging
- Primal/Dual Interpretation
- Nifty Tricks!

Subgradient method

algorithm:

$$t_k \text{ a sequence of step sizes}, x_1 = 0$$

$$\text{for } k = 1, 2, \ldots$$

$$g_k \in \partial f(x_k)$$

$$x_{k+1} = x_k - t_k g_k$$

Convergence

f convex and L-Lipschitz continuous

$$f(x_k^{\text{best}}) - f(x^*) \leq$$

$$f(\bar{x}_k) - f(x^*) \leq$$

$$\frac{2\|x_1 - x^*\|^2 + L^2 \sum_{i=1}^k t_i^2}{\sum_{i=1}^k t_i}$$

where

$$x_k^{\text{best}} = \underset{i \in \{1,2,\ldots,k\}}{\text{argmin}} \; f(x_i) \qquad \bar{x}_k = \frac{\sum_{i=1}^k t_i x_i}{\sum_{i=1}^k t_i}$$

Divergent step size

We must choose

$$\frac{2\|x_1 - x^*\|^2 + L^2 \sum_{i=1}^{k} t_i^2}{\sum_{i=1}^{k} t_i}$$

$$\sum_{i=1}^{\infty} t_i = \infty, \qquad \sum_{i=1}^{\infty} t_i^2 < \infty$$

to approach the minimum. *e.g.*,

$$t_k = \frac{1}{\sqrt{k}}, \qquad f(\bar{x}_k) - f(x^*) \in O\left(\frac{\log k}{\sqrt{k}}\right)$$

Divergent step size

We must choose

$$\frac{2\|x_1 - x^*\|^2 + L^2 \sum_{i=1}^{k} t_i^2}{\sum_{i=1}^{k} t_i}$$

$$\sum_{i=1}^{\infty} t_i = \infty, \qquad \sum_{i=1}^{\infty} t_i^2 < \infty$$

to approach the minimum. *e.g.*,

$$t_k = \frac{1}{\sqrt{k}}, \qquad f(\bar{x}_k) - f(x^*) \in O\left(\frac{\log k}{\sqrt{k}}\right)$$

Divergent step size

We must choose

$$\sum_{i=1}^{\infty} t_i = \infty, \qquad \sum_{i=1}^{\infty} t_i^2 < \infty$$

$$\frac{2\|x_1 - x^*\|^2 + L^2 \sum_{i=1}^{k} t_i^2}{\sum_{i=1}^{k} t_i}$$

to approach the minimum. *e.g.*,

$$t_k = \frac{1}{\sqrt{k}}, \qquad f(\bar{x}_k) - f(x^*) \in O\left(\frac{\log k}{\sqrt{k}}\right)$$

A closer look...

$$x_{k+1} = \operatorname*{argmin}_x f(x_k) + g_k \cdot (x - x_k) + \frac{\|x - x_k\|^2}{2t_k}$$

$$= \operatorname*{argmin}_x \sum_{i=1}^{k} \frac{t_i}{Z_k} \left[f(x_i) + g_i \cdot (x - x_i) \right] + \frac{\|x\|^2}{2Z_k}$$

where

$$Z_k = \sum_{i=1}^{k} t_i$$

The punch line

$$x_{k+1} = \operatorname*{argmin}_x \sum_{i=1}^{k} \frac{t_i}{Z_k} [f(x_i) + g_i \cdot (x - x_i)] + \frac{\|x\|^2}{2Z_k}$$

$$\bar{x}_k = \sum_{i=1}^{k} \frac{t_i}{Z_k} x_i$$

Why are new subgradients less important?

Dual Averaging

update:

$$x_{k+1} = \operatorname*{argmin}_x \frac{1}{k} \sum_{i=1}^{k} [f(x_i) + g_i \cdot (x - x_i)] + \frac{\mu_k \|x\|^2}{2k}$$

$$= \operatorname*{argmin}_x \hat{g}_k \cdot x + \frac{\mu_k \|x\|^2}{2k}$$

where

$$\hat{g}_k = \frac{1}{k} \sum_{i=1}^{k} g_i$$

Dual Averaging

update:

$$x_{k+1} = \operatorname*{argmin}_x \frac{1}{k} \sum_{i=1}^{k} [f(x_i) + g_i \cdot (x - x_i)] + \frac{\mu_k \|x\|^2}{2k}$$

$$= \operatorname*{argmin}_x \hat{g}_k \cdot x + \frac{\mu_k \|x\|^2}{2k}$$

"step size" control

where

$$\hat{g}_k = \frac{1}{k} \sum_{i=1}^{k} g_i$$

Dual Averaging

update:

$$x_{k+1} = \operatorname*{argmin}_x \frac{1}{k} \sum_{i=1}^{k} [f(x_i) + g_i \cdot (x - x_i)] + \frac{\mu_k \|x\|^2}{2k}$$

$$= \operatorname*{argmin}_x \hat{g}_k \cdot x + \frac{\mu_k \|x\|^2}{2k}$$

"step size" control

where

subgradients are equally important

$$\hat{g}_k = \frac{1}{k} \sum_{i=1}^{k} g_i$$

Dual Averaging

algorithm:

$$\mu_k \text{ a sequence of "step sizes"}, g_1 = 0$$

$$\text{for } k = 1, 2, \ldots$$

$$x_k = \frac{-k\hat{g}_k}{\mu_k}$$

$$g_{k+1} \in \partial f(x_k)$$

proof is not insightful, so let's skip it.

Convergence

f convex and L-Lipschitz continuous $\mu_k \in O\left(\sqrt{k}\right)$,

$$f(\hat{x}_k) - f(x^*) \leq \frac{0.5 + \sqrt{2k}}{k}\left(\|x_1 - x^*\|^2 + \frac{L^2}{2}\right)$$

$$f(\hat{x}_k) - f(x^*) \in O\left(\frac{1}{\sqrt{k}}\right)$$

where

$$\hat{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$$

Primal/Dual Problem

$$0 = \min_{x,g} f(x) + f^*(g) + D\|\text{-}g\|_*$$

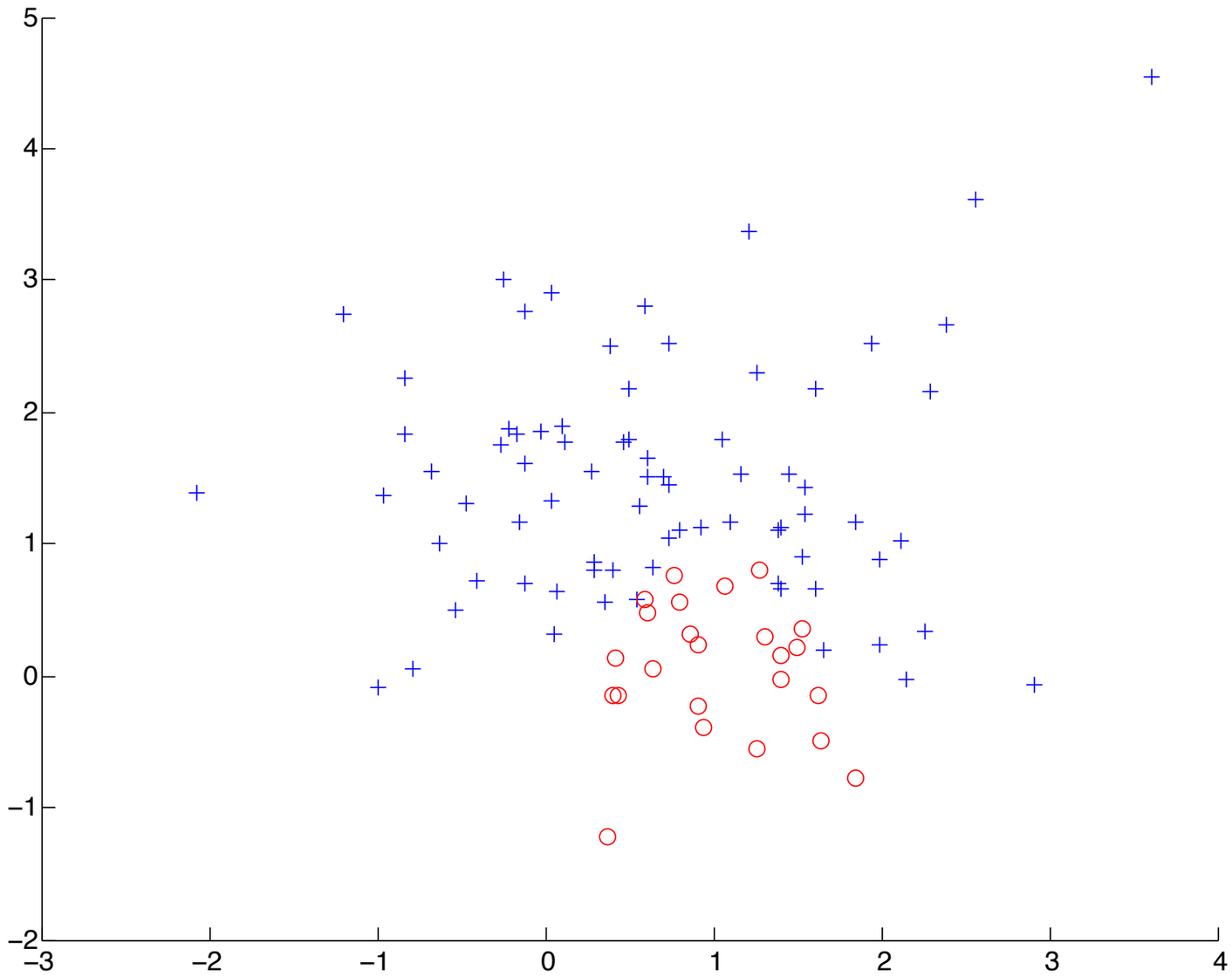
$$\text{when} \quad \|x^*\| \leq D$$

i.e., we can evaluate duality gap

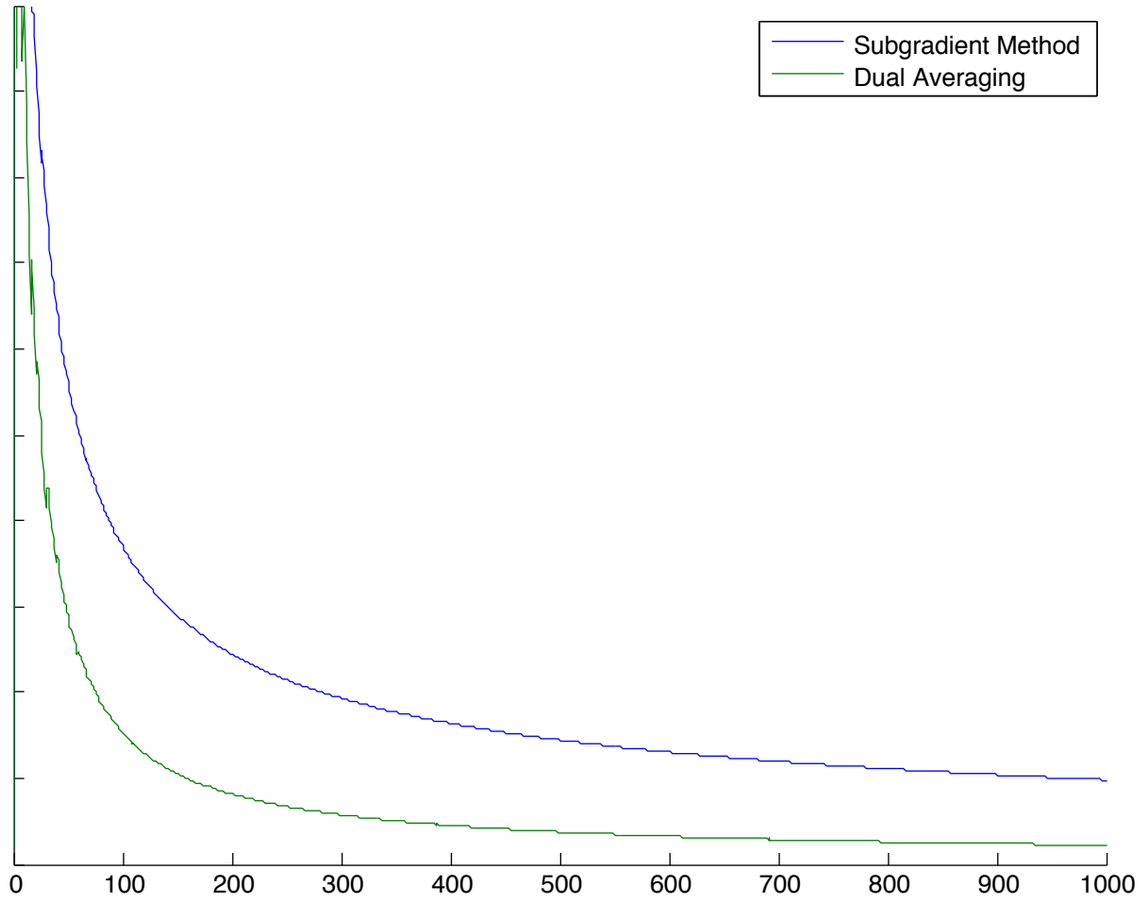
Should we use it?

- Can evaluate a duality gap*
- Beats SG method by log factor*
- Constants are often better*
  - "the [...] theory recommended the *worst possible* choice of parameters"
- Be careful when implementing update

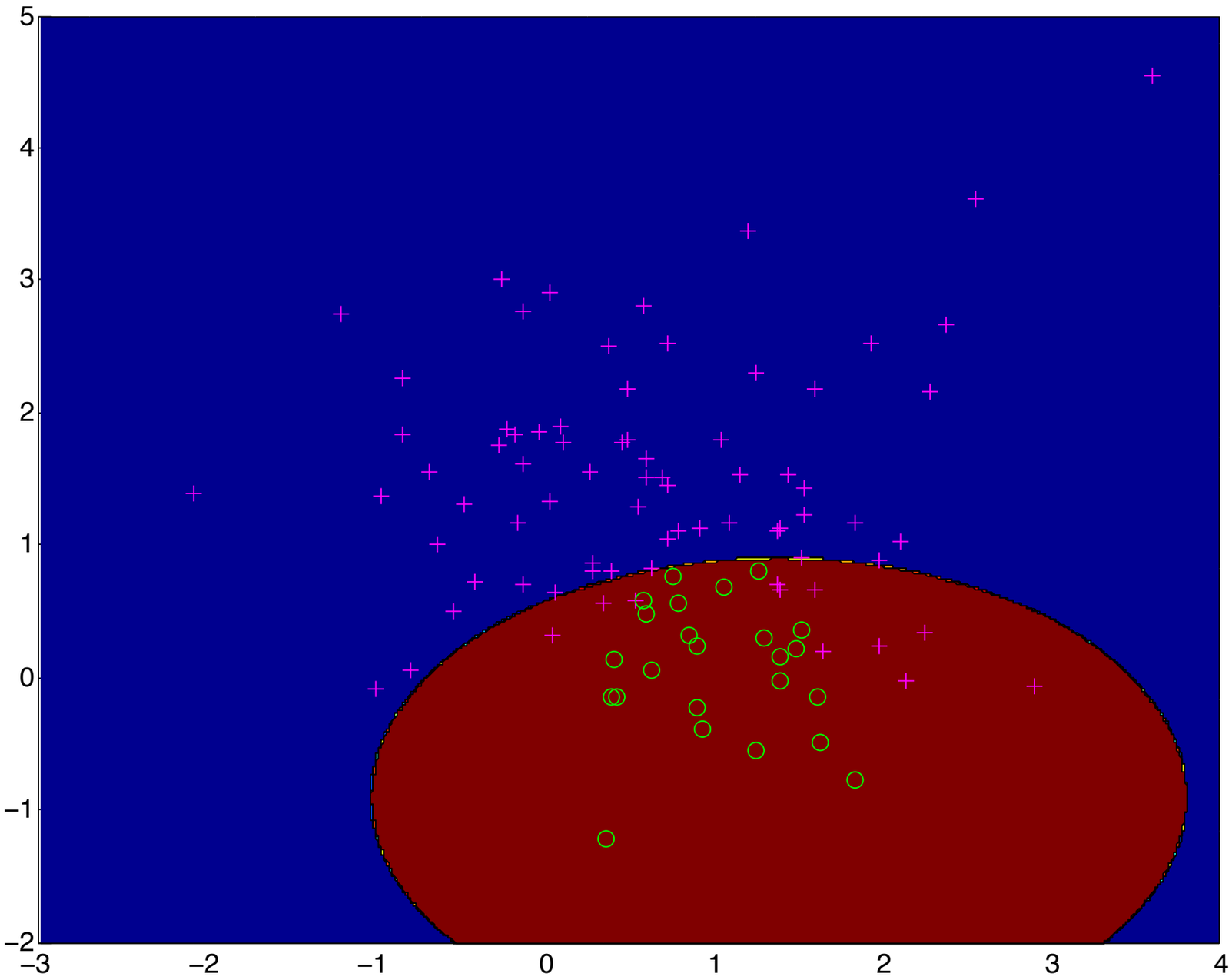
Demonstration: SVM



Convergence



Classifier



Strongly Convex

if $f(x)$ is strongly convex then

choosing $\mu_k = 1 + \log k$ gives convergence

$$f(\hat{x}_k) - f(x^*) \in O\left(\frac{\log k}{k}\right)$$

(can drop log by averaging differently)

Composite Objectives

$$f(x) = g(x) + h(x)$$

update:

$$x_{k+1} = \operatorname*{argmin}_x \frac{1}{k} \sum_{i=1}^{k} [f(x_i) + g_i \cdot (x - x_i)] + h(x) + \frac{\mu_k \|x\|^2}{2k}$$

$$= \operatorname*{argmin}_x \hat{g}_k \cdot x + h(x) + \frac{\mu_k \|x\|^2}{2k}$$

$$= \operatorname{prox}_{k/\mu_k} \left(\frac{k\hat{g}_k}{\mu_k} \right)$$

Bregman Divergences

update:

$$x_{k+1} = \operatorname*{argmin}_x \frac{1}{k} \sum_{i=1}^{k} [f(x_i) + g_i \cdot (x - x_i)] + \frac{\mu_k \Psi(x)}{k}$$

$$= \operatorname*{argmin}_x \hat{g}_k \cdot x + \frac{\mu_k \Psi(x)}{k}$$

where $\Psi(x)$ is a strongly convex function

Probability Simplex

e.g., $\Psi(x) = x \log x - 1 \cdot x$ for the unit simplex

$$x_{k+1} = \underset{x}{\operatorname{argmin}}\, \hat{g}_k \cdot x + \frac{\mu_k \left[x \log x - 1 \cdot x \right]}{k}$$

$$\text{subject to: } \sum_{i=1}^{n} x_i = 1, x \geq 0$$

$$x_{k+1} \propto \exp\left(\frac{-k \hat{g}_k}{\mu_k} \right)$$

Acceleration

$$x_1 = v_1 = \tilde{g}_1 = 0$$

for $k = 1, 2, \ldots$

$$y_k = (1 - \theta_k)x_k + \theta_k v_k$$

$$\tilde{g}_{k+1} = (1 - \theta_k)\tilde{g}_k + \theta_k \nabla f(y_k)$$

$$v_{k+1} = \operatorname*{argmin}_x \tilde{g}_{k+1} \cdot x + \frac{\mu_k \|x\|^2}{2}, \qquad \theta_k = \frac{2}{k+1}$$

$$x_{k+1} = (1 - \theta_k)x_k + \theta_k v_{k+1}$$

$$\mu_k = \frac{4(L + (k+1)^{3/2})}{k(k+1)}$$

convergence: $f(x_k) - f(x^*) \in O\left(\frac{L}{k^2}\right)$

Stochastic Objective

$$\min_x \mathbb{E}_\xi \left[f(x|\xi) \right]$$

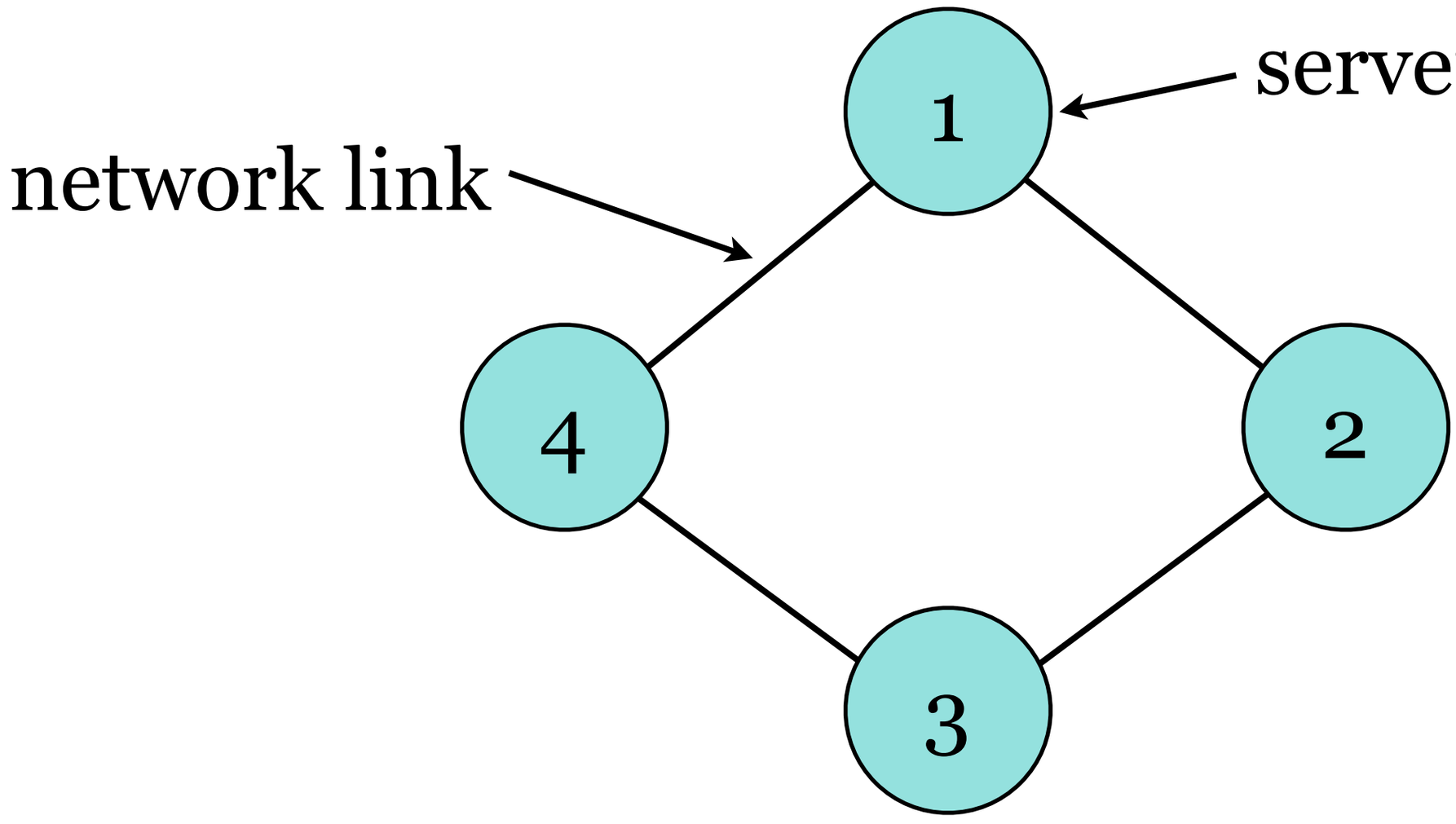
$$\mathbb{E}_\xi \left[g_k \right] \in \partial \mathbb{E}_\xi \left[f(x|\xi) \right]$$

$$\textbf{\textit{e.g.,}} \quad f(x) = \frac{1}{n} \sum_{i=1}^{n} (\theta_i \cdot x - y_i)^2$$

$$g_k = 2(\theta_{i_k} \cdot x_k - y_{i_k})\theta_{i_k}, \quad i_k \sim \text{Uniform}(\{1, 2, \ldots, n\})$$

Distributed Objective

$$\min_x f_1(x) + f_2(x) + f_3(x) + f_4(x)$$



Distributed DA

- Each server keeps its own primal and dual averages: $\hat{x}_k^i$, and $\tilde{g}_k^i$
- Every iteration, each server shares its dual average with all its neighbors: $\mathcal{N}(i)$
- All primal averages approach optimal!

Distributed DA

$$f(x) = \sum_{i=1}^{n} f_i(x) \qquad \tilde{g}^i_{k+1} = \frac{1}{k+1}\left[\sum_{j \in \mathcal{N}(i)} \frac{k\tilde{g}^j_k}{|\mathcal{N}(i)|} + g^i_k\right]$$

convergence: $f(\hat{x}^i_k) - f(x^*) \in O\left(\frac{1}{\sqrt{\gamma_G k}}\right)$ where

$\gamma_G = 1 - \sigma_2(G)$ is the spectral gap of the network

Graph Structures

- m-paths and cycles: $O\left(\frac{n}{m\sqrt{k}}\right)$
- m by m grid: $O\left(\frac{n^{1/4}}{m\sqrt{k}}\right)$
- expander graphs: $O\left(\frac{1}{\sqrt{k}}\right)$

Summary

- Y. Nesterov. Primal-dual subgradient methods for convex problems, 2005.
- L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization, 2011.
- J. Duchi, A. Agarwal, M. Wainwright. Dual averaging for distributed optimization: convergence analysis and network scaling, 2010.
- P. Tseng. On accelerated proximal gradient methods for convex-concave optimization, 2008.