On the Effectiveness of Richardson Extrapolation in Machine Learning

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

MAD+ Seminar - May 20, 2020

https://arxiv.org/pdf/2002.02835

https://francisbach.com/richardson-extrapolation/

Acceleration in numerical analysis

- **Principle**
  - Given asymptotic expansion in t around t_∞ (typically 0 or $+\infty$)

$$x_t = x_* + g_t + O(h_t),$$

    where $x_* \in \mathbb{R}^d$ is the desired output and $h_t = o(\|g_t\|)$
  - Combine iterates simply to obtain a sequence $y_t = x_* + O(h_t)$
  - Without the full knowledge of g_t

Acceleration in numerical analysis

- **Principle**
  - Given asymptotic expansion in t around t_∞ (typically 0 or $+\infty$)

$$x_t = x_* + g_t + O(h_t),$$

    where $x_* \in \mathbb{R}^d$ is the desired output and $h_t = o(\|g_t\|)$
  - Combine iterates simply to obtain a sequence $y_t = x_* + O(h_t)$
  - Without the full knowledge of g_t

- **Linear convergence** (exponential behavior)
  - Aitken's Δ^2 (Aitken, 1927), ε-algorithm (Wynn, 1956)
  - Anderson acceleration (Walker and Ni, 2011; Scieur et al., 2016)

Acceleration in numerical analysis

- **Principle**
  - Given asymptotic expansion in t around t_∞ (typically 0 or $+\infty$)

$$x_t = x_* + g_t + O(h_t),$$

    where $x_* \in \mathbb{R}^d$ is the desired output and $h_t = o(\|g_t\|)$
  - Combine iterates simply to obtain a sequence $y_t = x_* + O(h_t)$
  - Without the full knowledge of g_t

- **Linear convergence** (exponential behavior)
  - Aitken's Δ^2 (Aitken, 1927), ε-algorithm (Wynn, 1956)
  - Anderson acceleration (Walker and Ni, 2011; Scieur et al., 2016)

- **Sublinear convergence: Richardson extrapolation**

Richardson extrapolation (Richardson, 1911)

- **Sublinear convergence:** $x_t = x_* + t^\alpha \Delta + O(t^\beta)$
  - Linear combination $2x_t - x_{2^{1/\alpha}t}$

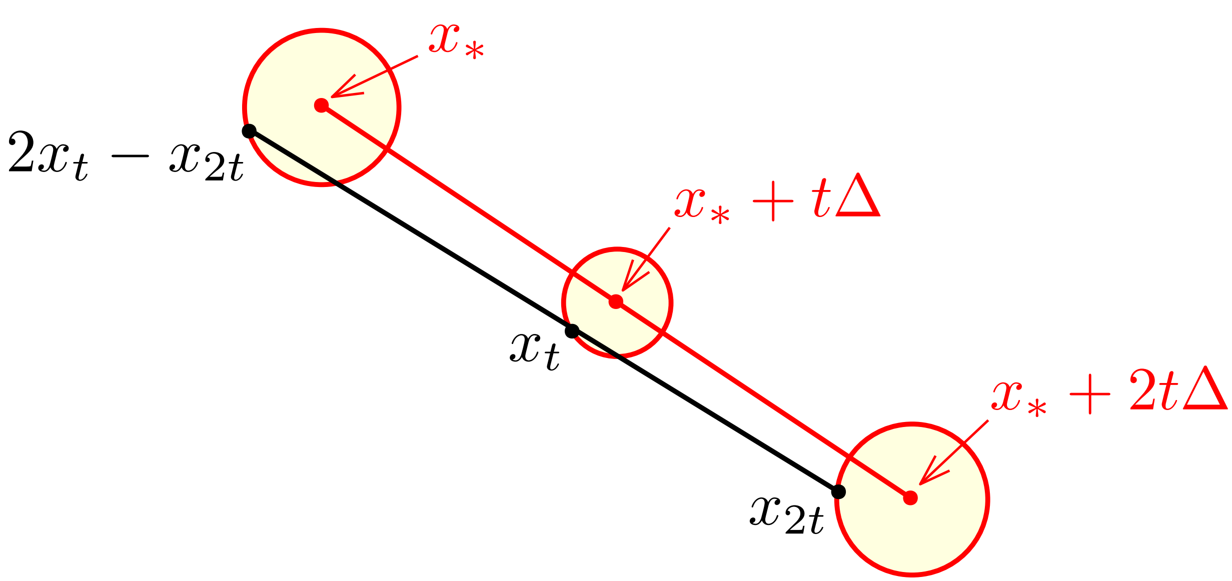
$$\begin{aligned} 2x_t - x_{2^{1/\alpha}t} &= 2(x_* + t^\alpha \Delta + O(t^\beta)) - (x_* + (2^{1/\alpha}t)^\alpha \Delta + O(t^\beta)) \\ &= x_* + O(t^\beta) \end{aligned}$$

Richardson extrapolation (Richardson, 1911)

- **Sublinear convergence:** $x_t = x_* + t^\alpha \Delta + O(t^\beta)$
  - Linear combination $2x_t - x_{2^{1/\alpha} t}$

$$
\begin{aligned}
2x_t - x_{2^{1/\alpha} t} &= 2(x_* + t^\alpha \Delta + O(t^\beta)) - (x_* + (2^{1/\alpha} t)^\alpha \Delta + O(t^\beta)) \\
&= x_* + O(t^\beta)
\end{aligned}
$$

  - Illustration with $t_\infty = 0$ and $\alpha = 1$, that is, $x_t = x_* + t\Delta + O(t^2)$



  - Typically used within integration methods (Richardson-Romberg)

Richardson extrapolation in machine learning

- **Iteration of an optimization algorithm:** $t = k \rightarrow +\infty$
  - Averaged gradient descent
  - Accelerated gradient descent
  - Frank-Wolfe algorithms

Richardson extrapolation in machine learning

- **Iteration of an optimization algorithm:** $t = k \to +\infty$
  - Averaged gradient descent
  - Accelerated gradient descent
  - Frank-Wolfe algorithms

- **Step-size of stochastic gradient descent:** $t = \gamma \to 0$

Richardson extrapolation in machine learning

- **Iteration of an optimization algorithm:** $t = k \to +\infty$
  - Averaged gradient descent
  - Accelerated gradient descent
  - Frank-Wolfe algorithms
- **Step-size of stochastic gradient descent:** $t = \gamma \to 0$
- **Regularization parameter:** $t = \lambda \to 0$
  - Nesterov smoothing
  - Ridge regression (not presented)

Richardson extrapolation in machine learning

- **Iteration of an optimization algorithm:** $t = k \to +\infty$
  - Averaged gradient descent
  - Accelerated gradient descent
  - Frank-Wolfe algorithms
- **Step-size of stochastic gradient descent:** $t = \gamma \to 0$
- **Regularization parameter:** $t = \lambda \to 0$
  - Nesterov smoothing
  - Ridge regression (not presented)
- **Requires asymptotic analysis**

Iteration of an optimization algorithm

- **Iterative algorithm** $x_k \in \mathbb{R}^d$, $k \geqslant 0$, with asymptotic expansion

$$x_k = x_* + \frac{1}{k}\Delta + O(1/k^2)$$

  – Extrapolation $x_k^{(1)} = 2x_k - x_{k/2}$ such that $x_k^{(1)} = x_* + O(1/k^2)$

Iteration of an optimization algorithm

- **Iterative algorithm** $x_k \in \mathbb{R}^d$, $k \geqslant 0$, with asymptotic expansion

$$x_k = x_* + \frac{1}{k}\Delta + O(1/k^2)$$

  – Extrapolation $x_k^{(1)} = 2x_k - x_{k/2}$ such that $x_k^{(1)} = x_* + O(1/k^2)$

- **When can we expect extrapolation to work?**
  – Having $\|x_k - x_*\|^2 = O(1/k^2)$ is not enough
  – Needs a specific asymptotic expansion

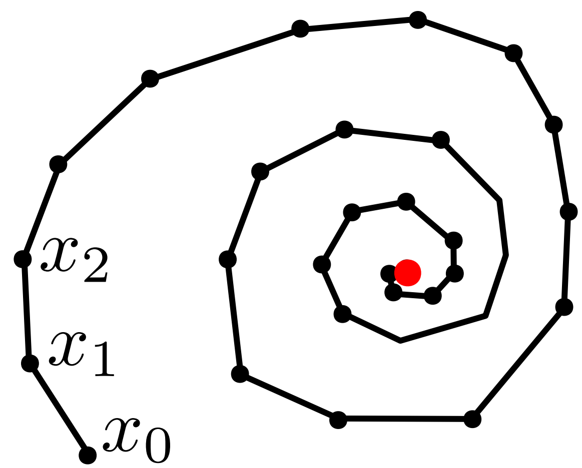
Iteration of an optimization algorithm

- **Iterative algorithm** $x_k \in \mathbb{R}^d$, $k \geqslant 0$, with asymptotic expansion

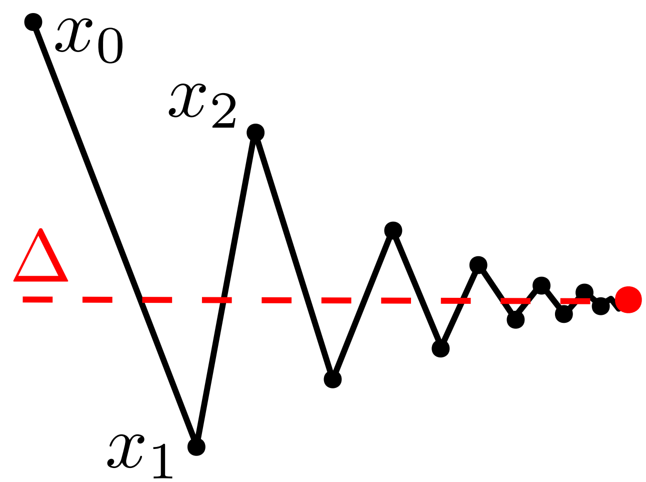
$$x_k = x_* + \frac{1}{k}\Delta + O(1/k^2)$$

  – Extrapolation $x_k^{(1)} = 2x_k - x_{k/2}$ such that $x_k^{(1)} = x_* + O(1/k^2)$

- **When can we expect extrapolation to work?**



oscillating convergence



non-oscillating convergence

Averaged gradient descent - I

- **Unconstrained minimization** $\min_{x \in \mathbb{R}^d} f(x)$
  - f convex, three-times differentiable
  - Hessian eigenvalues bounded
  - Unique minimizer $x_* \in \mathbb{R}^d$ such that $f''(x_*)$ is positive definite

Averaged gradient descent - I

- **Unconstrained minimization** $\min_{x \in \mathbb{R}^d} f(x)$
  - f convex, three-times differentiable
  - Hessian eigenvalues bounded
  - Unique minimizer $x_* \in \mathbb{R}^d$ such that $f''(x_*)$ is positive definite

- **Averaged gradient descent**

$$x_k = x_{k-1} - \gamma f'(x_{k-1}) \quad \text{and} \quad y_k = \frac{1}{k} \sum_{i=0}^{k-1} x_i$$

  - Averaging adds robustness to noise but forbids linear convergence
  - Polyak and Juditsky (1992); Nemirovski et al. (2009); Bach and Moulines (2011)

Averaged gradient descent - I

- **Unconstrained minimization** $\min_{x \in \mathbb{R}^d} f(x)$
  - f convex, three-times differentiable
  - Hessian eigenvalues bounded
  - Unique minimizer $x_* \in \mathbb{R}^d$ such that $f''(x_*)$ is positive definite

- **Averaged gradient descent**

$$x_k = x_{k-1} - \gamma f'(x_{k-1}) \quad \text{and} \quad y_k = \frac{1}{k} \sum_{i=0}^{k-1} x_i$$

  - Averaging adds robustness to noise but forbids linear convergence
  - Polyak and Juditsky (1992); Nemirovski et al. (2009); Bach and Moulines (2011)

- **Effect of Richardson extrapolation?**

Averaged gradient descent - II

$$x_k = x_{k-1} - \gamma f'(x_{k-1}) \quad \text{and} \quad y_k = \frac{1}{k}\sum_{i=0}^{k-1} x_i$$

- **Richardson extrapolation** (for k even)

$$y_k^{(1)} = 2y_k - y_{k/2} = \frac{2}{k}\sum_{i=0}^{k-1} x_i - \frac{2}{k}\sum_{i=0}^{k/2-1} x_i = \frac{2}{k}\sum_{i=k/2}^{k-1} x_i$$

  – Equivalent to tail-averaging (Jain et al., 2018)

Averaged gradient descent - II

$$x_k = x_{k-1} - \gamma f'(x_{k-1}) \quad \text{and} \quad y_k = \frac{1}{k}\sum_{i=0}^{k-1} x_i$$

- **Richardson extrapolation** (for k even)

$$y_k^{(1)} = 2y_k - y_{k/2} = \frac{2}{k}\sum_{i=0}^{k-1} x_i - \frac{2}{k}\sum_{i=0}^{k/2-1} x_i = \frac{2}{k}\sum_{i=k/2}^{k-1} x_i$$

  – Equivalent to tail-averaging (Jain et al., 2018)

- **Asymptotic expansion**: $y_k = x_* + \frac{1}{k}\Delta + O(\rho^k)$,

  where $\Delta = \sum_{i=0}^{\infty}(x_i - x_*)$ and $\rho \in (0,1)$

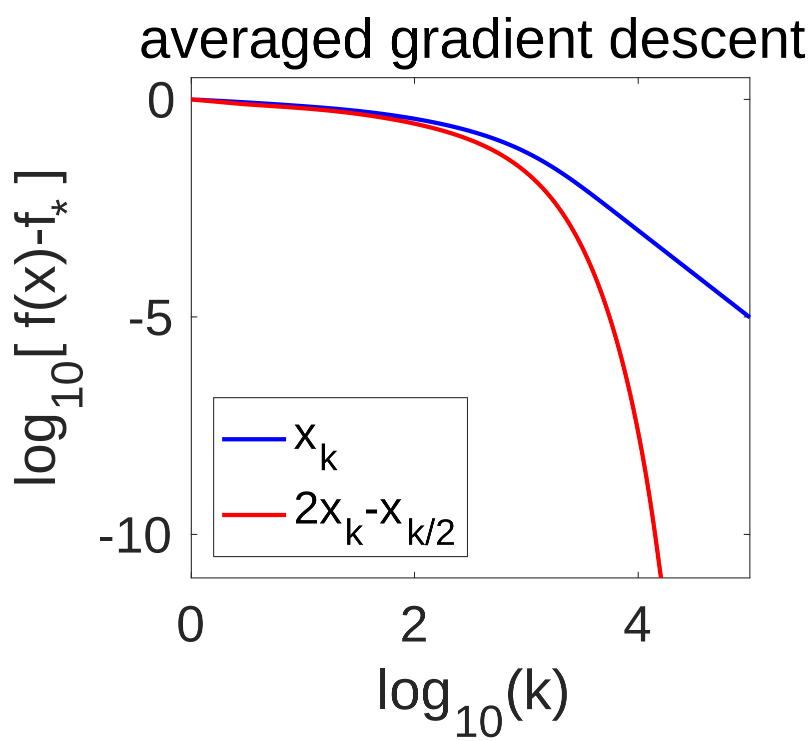
  – Richardson extrapolation restores linear convergence

Averaged gradient descent - III

- **Experiments on logistic regression**
  - Data $(a_i, b_i) \in \mathbb{R}^d \times \{-1, 1\}$, with $d = 400$ and $n = 4000$

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-b_i x^\top a_i))$$

  - Covariance matrix of inputs with eigenvalues $1/j$, $j = 1, \ldots, d$

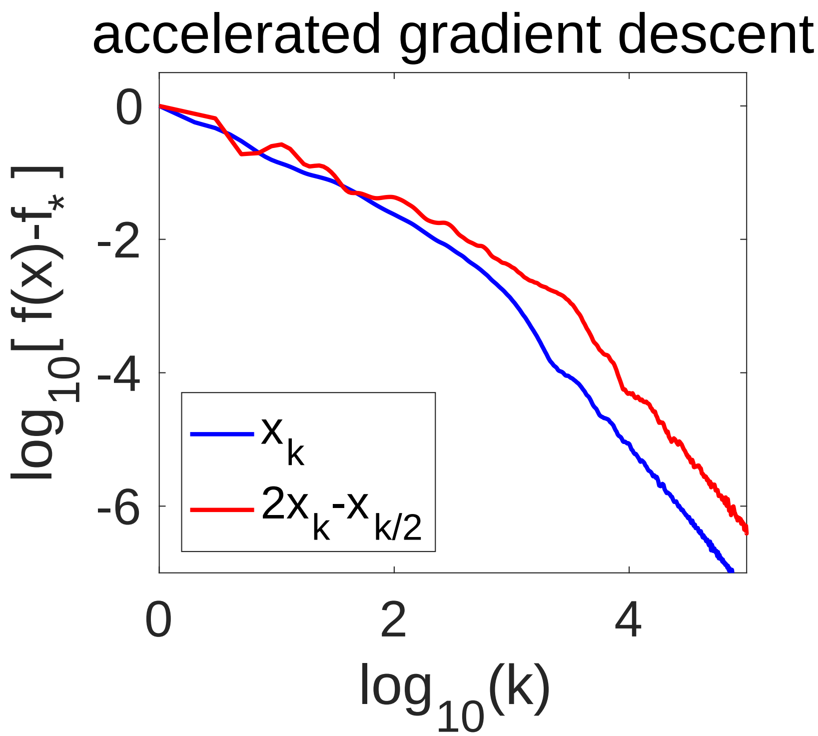


Accelerated gradient descent

- **Nesterov acceleration** (Nesterov, 1983)
  - Convergence in $O(1/k^2)$ instead of $O(1/k)$ for convex functions

Accelerated gradient descent

- **Nesterov acceleration** (Nesterov, 1983)
  - Convergence in $O(1/k^2)$ instead of $O(1/k)$ for convex functions
  - Iterates x_k oscillate around the optimum
    (see, e.g., Su et al., 2016; Flammarion and Bach, 2015)



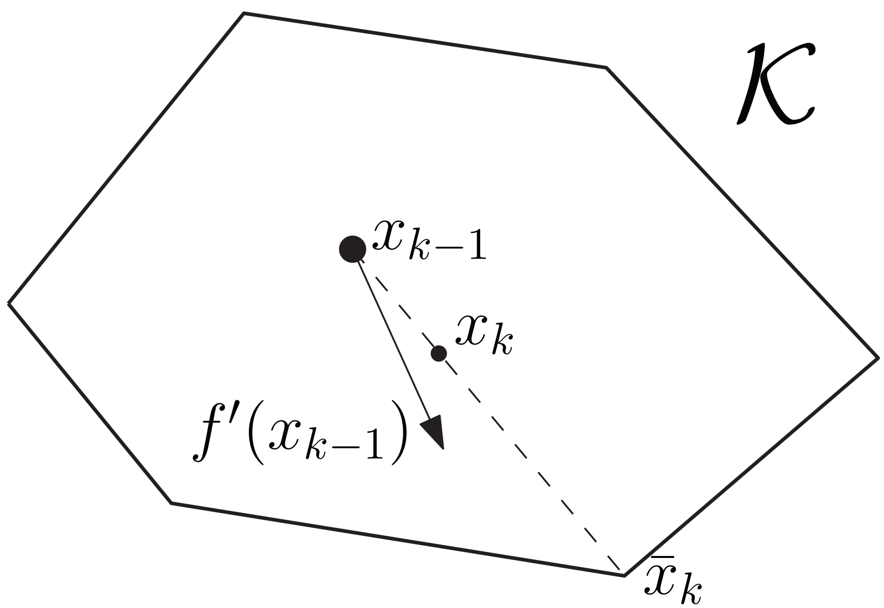
  - Richardson extrapolation is useless (but does not hurt)

Frank-Wolfe algorithms - I

- **Minimizing function f on a compact set $\mathcal{K}$**

$$\begin{aligned}\bar{x}_k &\in \arg\min_{x \in \mathcal{K}} \ f(x_{k-1}) + f'(x_{k-1})^\top (x - x_{k-1}) \\ x_k &= (1-\rho_k) x_{k-1} + \rho_k \bar{x}_k\end{aligned}$$

  - $\rho_k = 1/k$, $\rho = 2/(k+1)$ or with line search
  - Convergence rate: $f(x_k) - f(x_*) = O(1/k)$ or $O((\log k)/k)$
  - Dunn and Harshbarger (1978); Jaggi (2013)

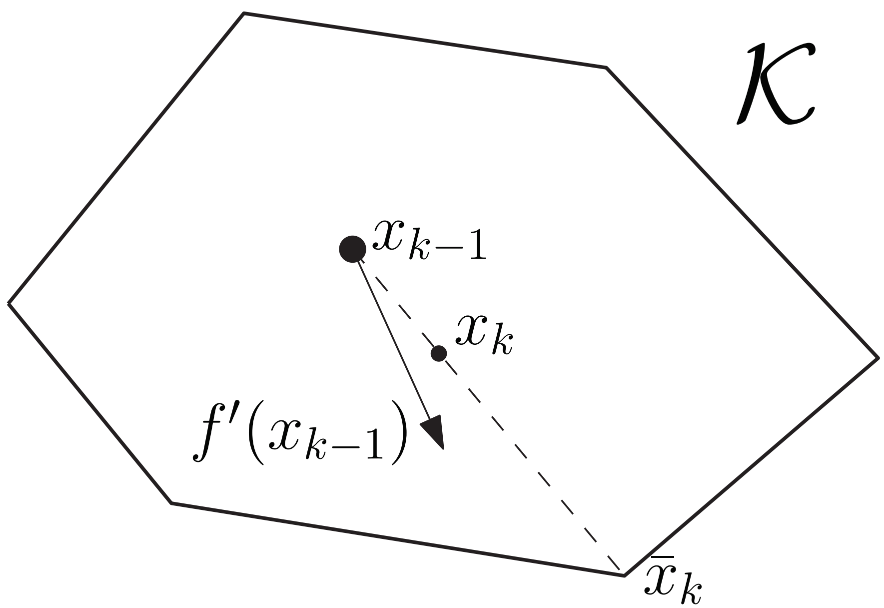


Frank-Wolfe algorithms - I

- **Minimizing function f on a compact set $\mathcal{K}$**

$$\begin{aligned} \bar{x}_k &\in \arg\min_{x \in \mathcal{K}} \ f(x_{k-1}) + f'(x_{k-1})^\top (x - x_{k-1}) \\ x_k &= (1-\rho_k) x_{k-1} + \rho_k \bar{x}_k \end{aligned}$$

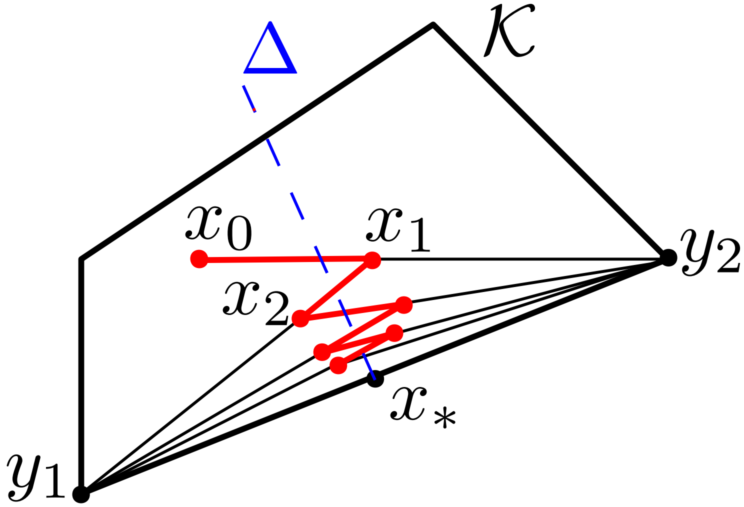
  - $\rho_k = 1/k$, $\rho = 2/(k+1)$ or with line search
  - Convergence rate: $f(x_k) - f(x_*) = O(1/k)$ or $O((\log k)/k)$
  - Dunn and Harshbarger (1978); Jaggi (2013)



- **Effect of Richardson extrapolation?**

Frank-Wolfe algorithms - II

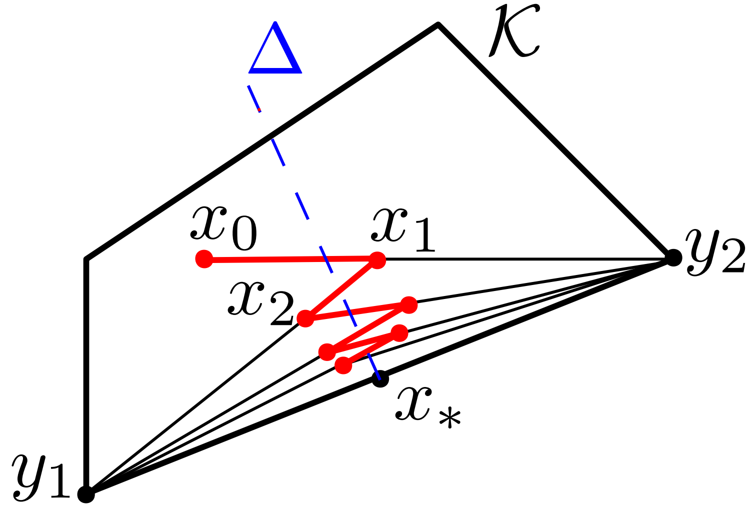
- **Asssumptions:** $\mathcal{K}$ polytope + "constraint qualification"
- **Step-size** $\rho_k = 1/k$
  - Asymptotic expansion: $x_k = x_* + \frac{1}{k}\Delta_1 + O(1/k^2)$
  - With Δ_1 orthogonal the facet of x_* in $\mathcal{K}$



Frank-Wolfe algorithms - II

- **Asssumptions:** $\mathcal{K}$ polytope + "constraint qualification"

- **Step-size** $\rho_k = 1/k$
  - Asymptotic expansion: $x_k = x_* + \dfrac{1}{k}\Delta_1 + O(1/k^2)$
  - With Δ_1 orthogonal the facet of x_* in $\mathcal{K}$



  - Function values: $f(x_k) - f(x_*) = \frac{1}{k}\Delta_1^\top f'(x_*) + O(1/k^2)$
  - Richardson: $f(2x_k - x_{k/2}) - f(x_*) = O(1/k^2)$
  - Richardson extrapolation transforms $O(1/k)$ to $O(1/k^2)$

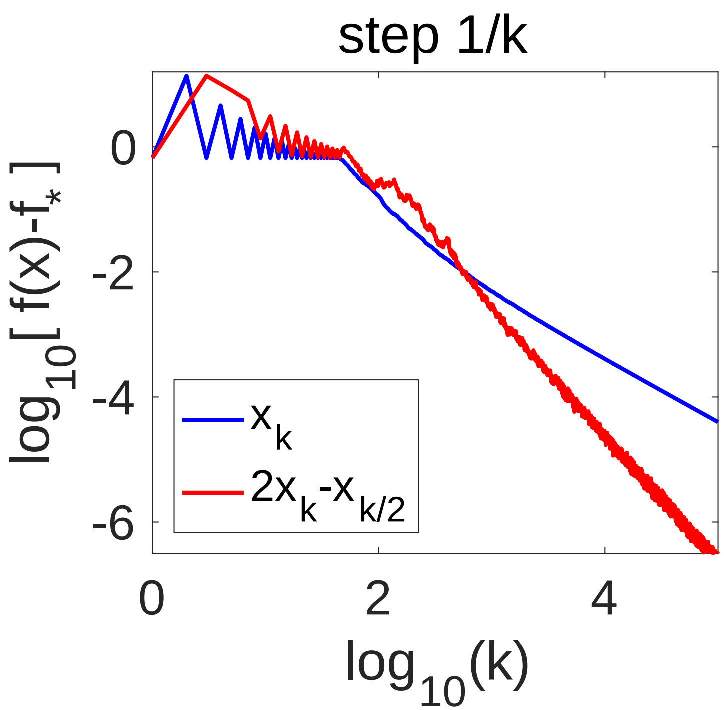
Frank-Wolfe algorithms - III

- **Step-size** $\rho_k = 1/k$

- **Experiments on constrained logistic regression**

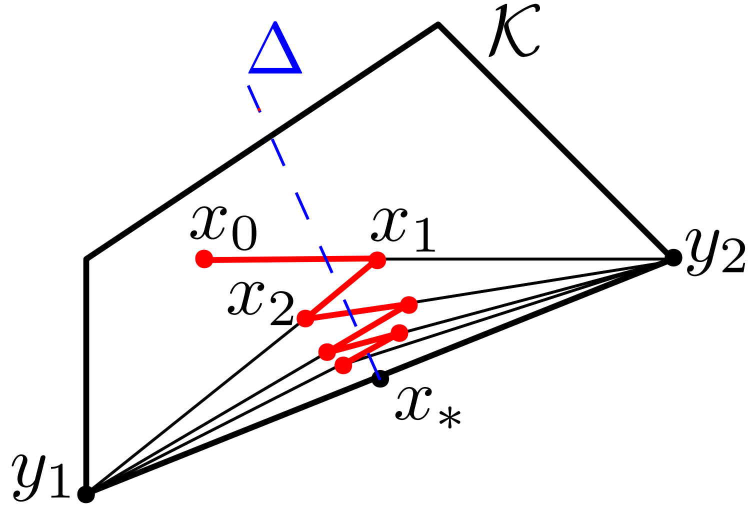
  – Data $(a_i, b_i) \in \mathbb{R}^d \times \{-1, 1\}$, with $d = 400$ and $n = 400$

$$\min_{\|x\|_1 \leqslant c} \quad \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-b_i x^\top a_i))$$



Frank-Wolfe algorithms - IV

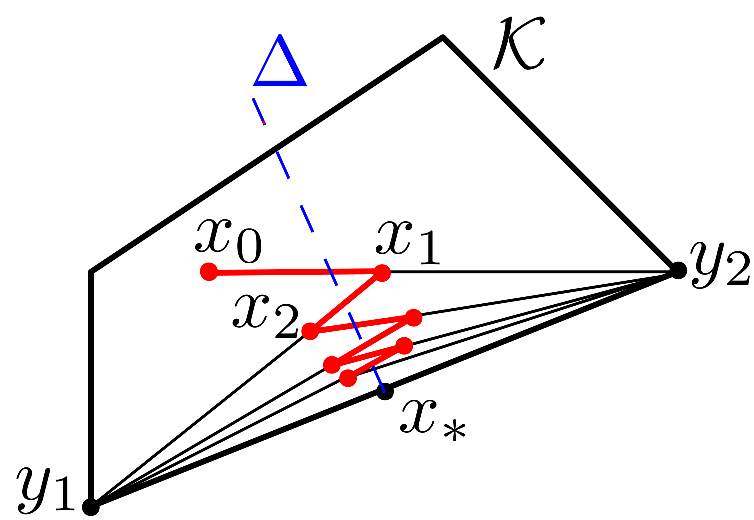
- **Asssumptions:** $\mathcal{K}$ polytope + “constraint qualification”
- **Step-size** $\rho_k = 2/(k+1)$
  - Asymptotic expansion: $x_k = x_* + \frac{1}{k(k+1)}\Delta_2 + O(1/k^2)$
  - With Δ_2 orthogonal the facet of x_* in $\mathcal{K}$



Frank-Wolfe algorithms - IV

- **Asssumptions:** $\mathcal{K}$ polytope + "constraint qualification"

- **Step-size** $\rho_k = 2/(k+1)$
  - Asymptotic expansion: $x_k = x_* + \frac{1}{k(k+1)}\Delta_2 + O(1/k^2)$
  - With Δ_2 orthogonal the facet of x_* in $\mathcal{K}$



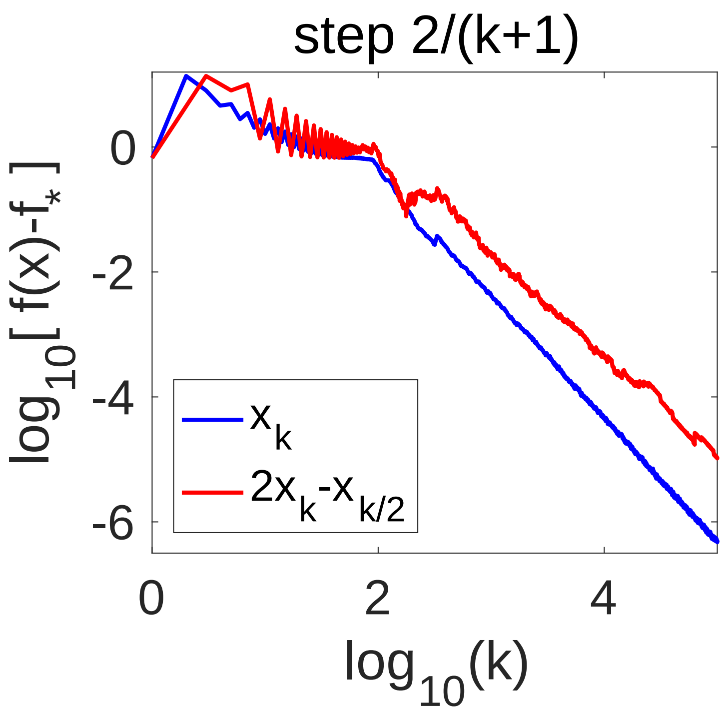
  - Function values: $f(x_k) - f(x_*) = O(1/k^2)$
  - Richardson: $f(2x_k - x_{k/2}) - f(x_*) = O(1/k^2)$
  - Richardson extrapolation is useless (but does not hurt)

Frank-Wolfe algorithms - V

- **Step-size** $\rho_k = 2/(k+1)$

- **Experiments on constrained logistic regression**
  - Data $(a_i, b_i) \in \mathbb{R}^d \times \{-1, 1\}$, with $d = 400$ and $n = 400$

$$\min_{\|x\|_1 \leqslant c} \quad \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-b_i x^\top a_i))$$

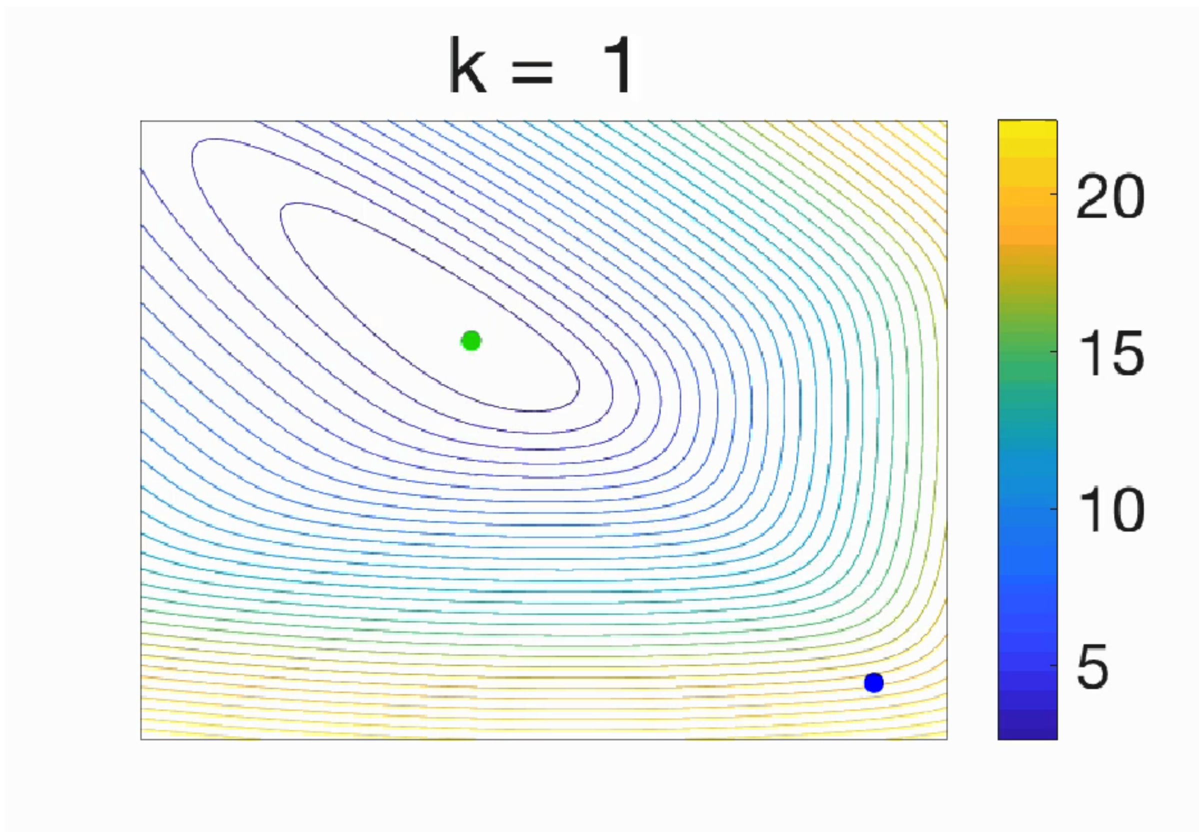


Step-size of stochastic gradient descent - I

- **Averaged SGD**, with stochastic gradients $g'(x_{k-1}, z_k)$

$$x_k = x_{k-1} - \gamma g'(x_{k-1}, z_k) \quad \text{and} \quad y_k = \frac{1}{k}\sum_{i=0}^{k-1} x_i$$

  - with expectation $\mathbb{E}_{z_k} g'(x_{k-1}, z_k) = f'(x_{k-1})$
  - y_k converges to $y_*^{(\gamma)} \neq x_* = \arg\min f$ (Dieuleveut et al., 2017)



Step-size of stochastic gradient descent - I

- **Averaged SGD**, with stochastic gradients $g'(x_{k-1}, z_k)$

$$x_k = x_{k-1} - \gamma g'(x_{k-1}, z_k) \quad \text{and} \quad y_k = \frac{1}{k}\sum_{i=0}^{k-1} x_i$$

  - with expectation $\mathbb{E}_{z_k} g'(x_{k-1}, z_k) = f'(x_{k-1})$
  - y_k converges to $y_*^{(\gamma)} \neq x_* = \arg\min f$ (Dieuleveut et al., 2017)

- **Asymptotic expansion:** $y_*^{(\gamma)} = x_* + \gamma\Delta + O(\gamma^2)$
  - Richardson extrapolation $2y_n^{(\gamma)} - y_n^{(2\gamma)}$ converges to

$$2y_*^{(\gamma)} - y_*^{(2\gamma)} = x_* + O(\gamma^2)$$

Step-size of stochastic gradient descent - I

- **Averaged SGD**, with stochastic gradients $g'(x_{k-1}, z_k)$

$$x_k = x_{k-1} - \gamma g'(x_{k-1}, z_k) \quad \text{and} \quad y_k = \frac{1}{k} \sum_{i=0}^{k-1} x_i$$

  – with expectation $\mathbb{E}_{z_k} g'(x_{k-1}, z_k) = f'(x_{k-1})$
  – y_k converges to $y_*^{(\gamma)} \neq x_* = \arg\min f$ (Dieuleveut et al., 2017)

- **Asymptotic expansion:** $y_*^{(\gamma)} = x_* + \gamma \Delta + O(\gamma^2)$
  – Richardson extrapolation $2y_n^{(\gamma)} - y_n^{(2\gamma)}$ converges to

$$2y_*^{(\gamma)} - y_*^{(2\gamma)} = x_* + O(\gamma^2)$$

  – Higher-order extrapolation $3y_n^{(\gamma)} - 3y_n^{(2\gamma)} + y_n^{(3\gamma)}$ removes the term in γ^2 and approaches x_* with rate $O(\gamma^3)$

Step-size of stochastic gradient descent - I

- **Averaged SGD**, with stochastic gradients $g'(x_{k-1}, z_k)$

$$x_k = x_{k-1} - \gamma g'(x_{k-1}, z_k) \quad \text{and} \quad y_k = \frac{1}{k}\sum_{i=0}^{k-1} x_i$$

  – with expectation $\mathbb{E}_{z_k} g'(x_{k-1}, z_k) = f'(x_{k-1})$
  – y_k converges to $y_*^{(\gamma)} \neq x_* = \arg\min f$ (Dieuleveut et al., 2017)

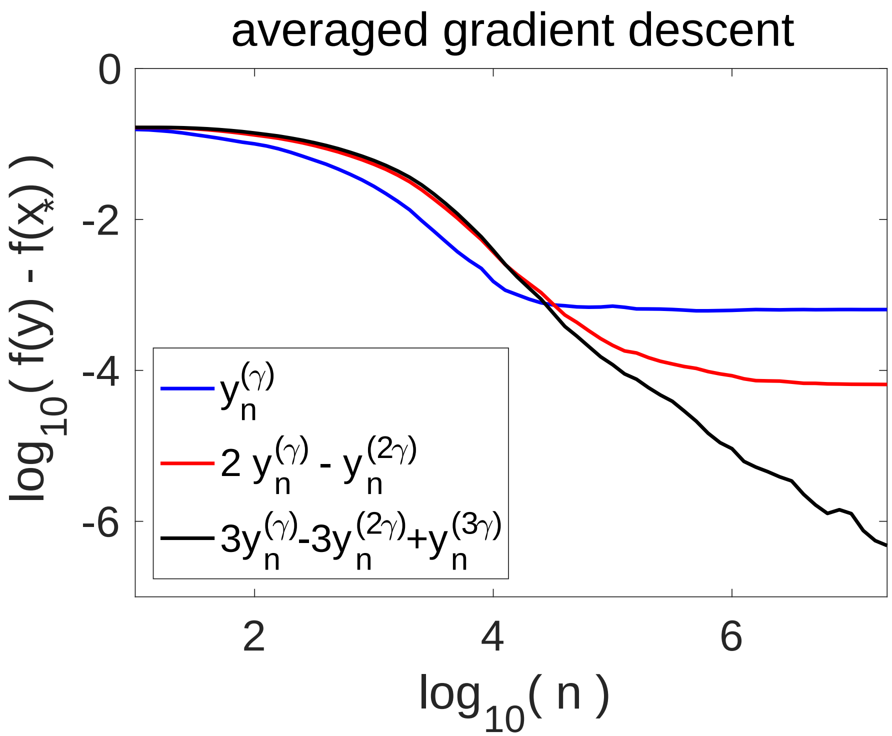
- **Asymptotic expansion:** $y_*^{(\gamma)} = x_* + \gamma\Delta + O(\gamma^2)$
  – Richardson extrapolation $2y_n^{(\gamma)} - y_n^{(2\gamma)}$ converges to

$$2y_*^{(\gamma)} - y_*^{(2\gamma)} = x_* + O(\gamma^2)$$

  – Higher-order extrapolation $3y_n^{(\gamma)} - 3y_n^{(2\gamma)} + y_n^{(3\gamma)}$ removes the term in γ^2 and approaches x_* with rate $O(\gamma^3)$
  – Can go up to order m...

Step-size of stochastic gradient descent - II

- **Experiments on logistic regression** in dimension 20
  - Dieuleveut, Durmus, and Bach (2017)



  - See also Durmus, Simsekli, Moulines, Badeau, and Richard (2016)

Nesterov smoothing - I

- **Composite problem**: minimize $f = h + g$
  - With h smooth and g non-smooth
  - Structured prediction, or sparsity-inducing norms

Nesterov smoothing - I

- **Composite problem**: minimize $f = h + g$
  - With h smooth and g non-smooth
  - Structured prediction, or sparsity-inducing norms

- **Nesterov smoothing** (Nesterov, 2005): replace g by g_λ
  - With g_λ is $(1/\lambda)$-smooth, and $\|g - g_\lambda\|_\infty = O(\lambda)$
  - Typically done by inf-convolution with a $(1/\lambda)$-smooth function
  - Example: smooth $\max\{x, y\}$ by $\lambda \log(\exp(x/\lambda) + \exp(y/\lambda))$

Nesterov smoothing - I

- **Composite problem**: minimize $f = h + g$
  - With h smooth and g non-smooth
  - Structured prediction, or sparsity-inducing norms

- **Nesterov smoothing** (Nesterov, 2005): replace g by g_λ
  - With g_λ is $(1/\lambda)$-smooth, and $\|g - g_\lambda\|_\infty = O(\lambda)$
  - Typically done by inf-convolution with a $(1/\lambda)$-smooth function
  - Example: smooth $\max\{x, y\}$ by $\lambda \log(\exp(x/\lambda) + \exp(y/\lambda))$

- **Optimization of $h + g_\lambda$ by accelerated gradient descent**
  - Error rate of $O\big(\lambda + 1/(\lambda k^2)\big)$
  - With $\lambda \propto 1/k$, rate of $O(1/k)$
  - Better than subgradient method in $O(1/\sqrt{k})$

Nesterov smoothing - II

- **Assumptions**: (1) polyhedral function g
  (2) smoothing by entropic or quadratic dual penalty

- **Asymptotic expansion**
  - If x_λ is the minimizer of $h + g_\lambda$
  - If x_* the global minimizer of $f = h + g$

$$x_\lambda = x_* + \lambda\Delta + O(\lambda^2)$$

Nesterov smoothing - II

- **Assumptions**: (1) polyhedral function g
  (2) smoothing by entropic or quadratic dual penalty

- **Asymptotic expansion**
  - If x_λ is the minimizer of $h + g_\lambda$
  - If x_* the global minimizer of $f = h + g$

$$x_\lambda = x_* + \lambda\Delta + O(\lambda^2)$$

  - Then $x_\lambda^{(1)} = 2x_\lambda - x_{2\lambda} = x_* + O(\lambda^2)$ and $f(x_\lambda^{(1)}) = f(x_*) + O(\lambda^2)$
  - Error rate of $O\big(\lambda^2 + 1/(\lambda k^2)\big)$
  - With $\lambda \propto k^{-2/3}$, overall convergence rate of $k^{-4/3}$

Nesterov smoothing - II

- **Assumptions**: (1) polyhedral function g
  (2) smoothing by entropic or quadratic dual penalty

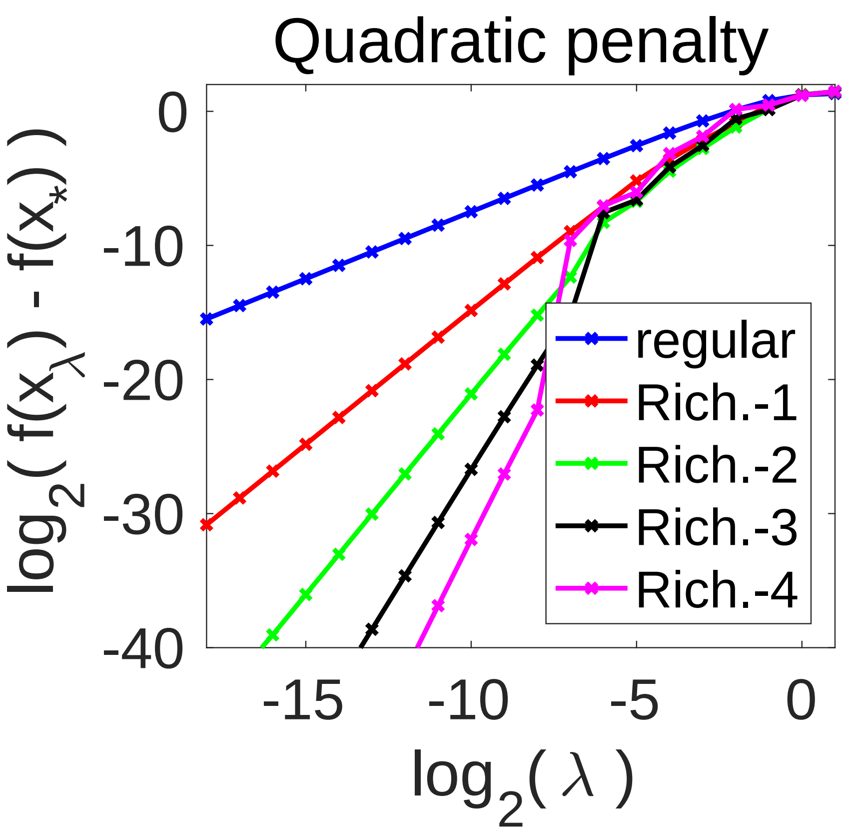
- **Asymptotic expansion**
  - If x_λ is the minimizer of $h + g_\lambda$
  - If x_* the global minimizer of $f = h + g$

$$x_\lambda = x_* + \lambda\Delta + O(\lambda^2)$$

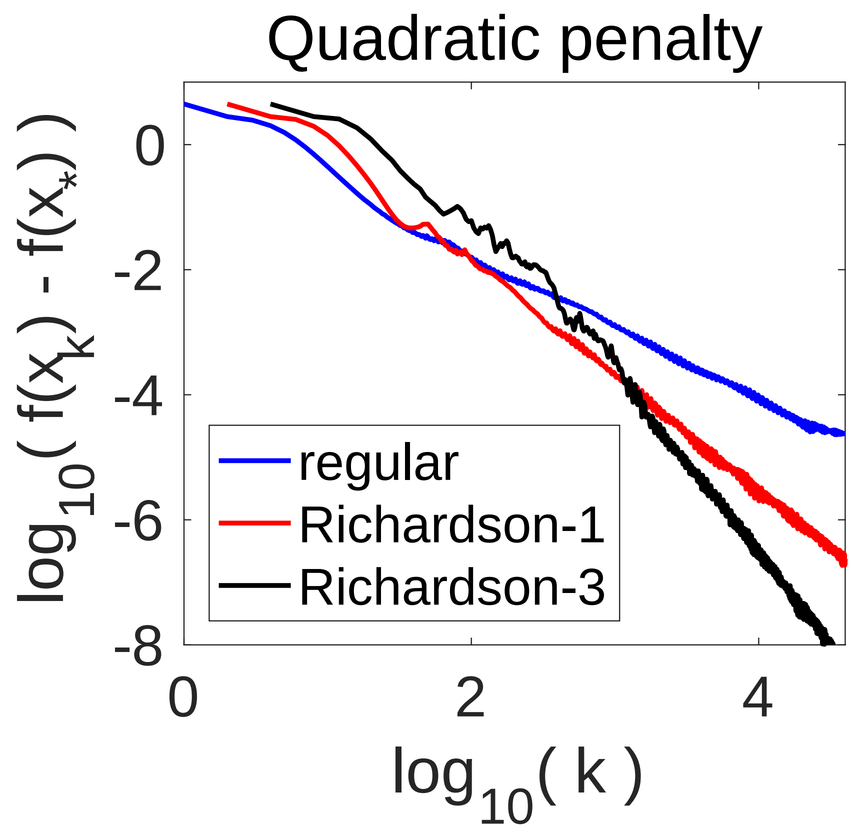
  - Then $x_\lambda^{(1)} = 2x_\lambda - x_{2\lambda} = x_* + O(\lambda^2)$ and $f(x_\lambda^{(1)}) = f(x_*) + O(\lambda^2)$
  - Error rate of $O\big(\lambda^2 + 1/(\lambda k^2)\big)$
  - With $\lambda \propto k^{-2/3}$, overall convergence rate of $k^{-4/3}$
  - High-order expansions have rate $O(k^{-2(m+1)/(m+2)})$

Nesterov smoothing - III

- **Experiments on penalized Lasso problem**



Study of x_λ



Study of x_k

Richardson extrapolation in machine learning

- **Iteration of an optimization algorithm:** $t = k \to +\infty$
  - Averaged gradient descent
  - Accelerated gradient descent
  - Frank-Wolfe algorithms
- **Step-size of stochastic gradient descent:** $t = \gamma \to 0$
- **Regularization parameter:** $t = \lambda \to 0$
  - Nesterov smoothing
  - Ridge regression (not presented)
- **Requires asymptotic analysis**

Richardson extrapolation in machine learning

- **Iteration of an optimization algorithm:** $t = k \to +\infty$
  - Averaged gradient descent
  - Accelerated gradient descent
  - Frank-Wolfe algorithms
- **Step-size of stochastic gradient descent:** $t = \gamma \to 0$
- **Regularization parameter:** $t = \lambda \to 0$
  - Nesterov smoothing
  - Ridge regression (not presented)
- **Requires asymptotic analysis**
- **Other problems?**

References

Alexander Craig Aitken. Xxv.on bernoulli's numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh*, 46:289–305, 1927.

F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. Technical Report 1707.06386, arXiv, 2017.

J. C. Dunn and S. Harshbarger. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432–444, 1978.

Alain Durmus, Umut Simsekli, Eric Moulines, Roland Badeau, and Gaël Richard. Stochastic gradient Richardson-Romberg Markov chain Monte Carlo. In *Advances in Neural Information Processing Systems*, 2016.

N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. *arXiv preprint arXiv:1504.01577*, 2015.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, 2013.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18(223):1–42, 2018.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Y. Nesterov. A method for solving a convex programming problem with rate of convergence $O(1/k^2)$. *Soviet Math. Doklady*, 269(3):543–547, 1983.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1): 127–152, 2005.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Lewis Fry Richardson. The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Philosophical Transactions of the Royal Society of London. Series A*, 210(459-470):307–357, 1911.

D. Scieur, A. d'Aspremont, and F. Bach. Regularized nonlinear acceleration. In *Advances In Neural Information Processing Systems*, 2016.

Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *Journal of Machine Learning Research*, 17(1): 5312–5354, 2016.

Homer F. Walker and Peng Ni. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49(4):1715–1735, 2011.

Peter Wynn. On a device for computing the $e_m(S_n)$ transformation. *Mathematical Tables and Other Aids to Computation*, pages 91–96, 1956.