Stochastic Optimization: First order method

† ‡ Taiji Suzuki

†Tokyo Institute of Technology
Graduate School of Information Science and Engineering
Department of Mathematical and Computing Sciences
‡JST, PRESTO

Intensive course @ Nagoya University

Outline

1. First order method
   - Proximal gradient descent
   - Nesterov's acceleration and optimal convergence

Outline

1. First order method
   - Proximal gradient descent
   - Nesterov's acceleration and optimal convergence

Regularized learning problem

Lasso:

$$\min_{x \in \mathbb{R}^p} \quad \frac{1}{n} \sum_{i=1}^{n} (y_i - z_i^\top x)^2 + \underbrace{\|x\|_1}_{\text{regularization}} .$$

Regularized learning problem

Lasso:

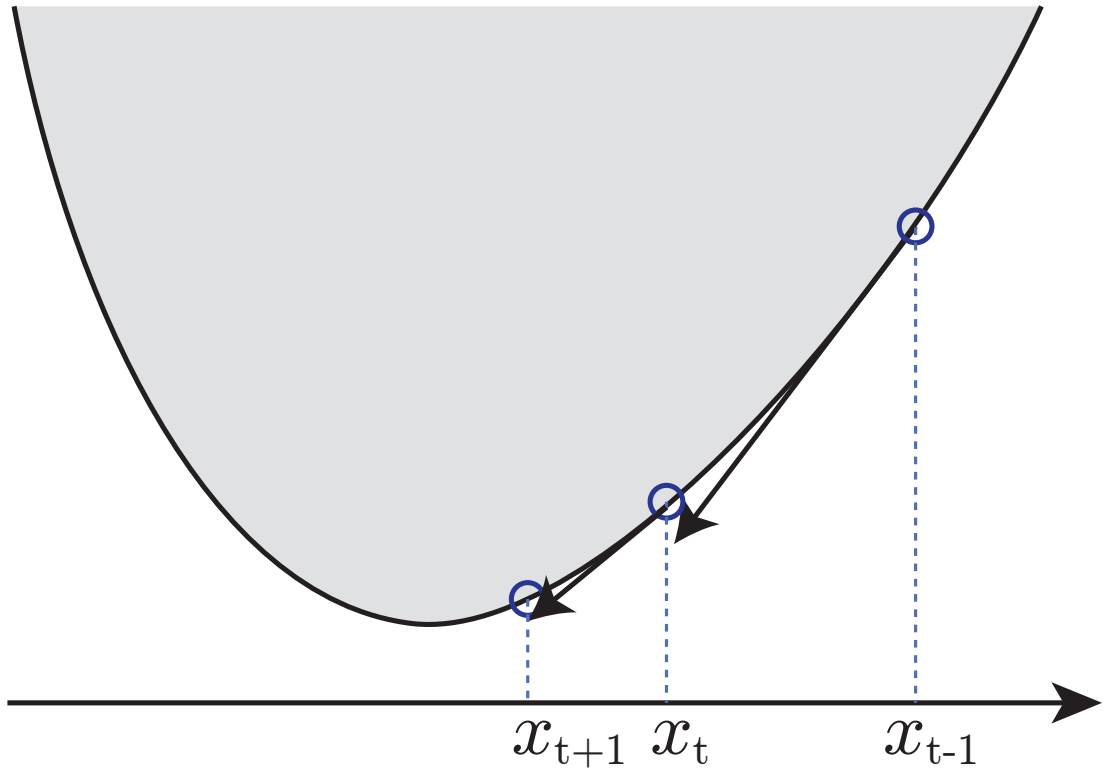
$$\min_{x \in \mathbb{R}^p} \quad \frac{1}{n} \sum_{i=1}^{n} (y_i - z_i^\top x)^2 + \underbrace{\|x\|_1}_{\text{regularization}} .$$

General regularized learning problem:

$$\min_{x \in \mathbb{R}^p} \quad \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, x) \; + \; \psi(x).$$

Difficulty: Sparsity inducing regularization is usually **non-smooth**.

First order optimization



- Optimization methods that use only the function value $f(x)$ and the first order gradient $g \in \partial f(x)$.
- Computation per iteration is light, and suited for high dimensional problems.
- Newton method is a second order method.

Outline

1. First order method
   - Proximal gradient descent
   - Nesterov's acceleration and optimal convergence

Gradient descent

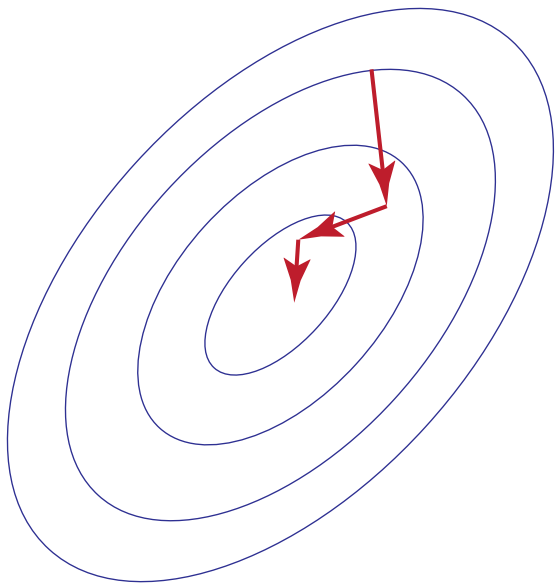
Let $f(x) = \sum_{i=1}^{n} \ell(z_i, x)$.

$$\min_x \quad f(x).$$

Subgradient method

Differentiable $f(x)$:

$$x_t = x_{t-1} - \eta_t \nabla f(x_{t-1}).$$



Gradient descent

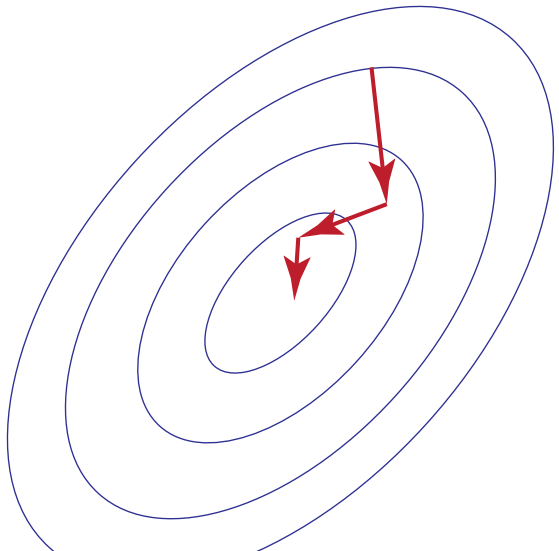
Let $f(x) = \sum_{i=1}^{n} \ell(z_i, x)$.

$$\min_x \quad f(x).$$

Subgradient method

Subdifferentiable $f(x)$:

$$g_t \in \partial f(x_{t-1}),$$
$$x_t = x_{t-1} - \eta_t g_t.$$



Gradient descent

Let $f(x) = \sum_{i=1}^{n} \ell(z_i, x)$.

$$\min_x \quad f(x).$$

Subgradient method (equivalent formula)

Subdifferentiable $f(x)$:

$$x_t = \operatorname*{argmin}_x \left\{ \langle x, g_t \rangle + \frac{1}{2\eta_t} \|x - x_{t-1}\|^2 \right\},$$

where $g_t \in \partial f(x_{t-1})$.

Proximal point algorithm:

$$x_t = \operatorname*{argmin}_x \left\{ f(x) + \frac{1}{2\eta_t} \|x - x_{t-1}\|^2 \right\}.$$

- $f(x_t) \to$ optimum for any convex f and $\eta_t = \eta > 0$ (**?**).
- If $f(x)$ is strongly convex: $f(x_t) - f(x^*) \leq \frac{1}{2\eta} \left(\frac{1}{1+\sigma\eta} \right)^{t-1} \|x_0 - x^*\|^2$.

Proximal gradient descent

Let $f(x) = \sum_{i=1}^{n} \ell(z_i, x)$.

$$\min_x \quad f(x) + \psi(x).$$

Proximal gradient descent

$$x_t = \operatorname*{argmin}_x \left\{ \langle x, g_t \rangle + \psi(x) + \frac{1}{2\eta_t} \|x - x_{t-1}\|^2 \right\}$$
$$= \operatorname*{argmin}_x \left\{ \eta_t \psi(x) + \frac{1}{2} \|x - (x_{t-1} - \eta_t g_t)\|^2 \right\}$$

where $g_t \in \partial f(x_{t-1})$.

The update rule is given by **proximal mapping**:

$$\mathrm{prox}(q|\tilde{\psi}) = \operatorname*{argmin}_x \left\{ \tilde{\psi}(x) + \frac{1}{2} \|x - q\|^2 \right\}$$

$\rightarrow$ By using the proximal mapping, we can avoid bad properties (e.g., non-smoothness) of ψ.

Example

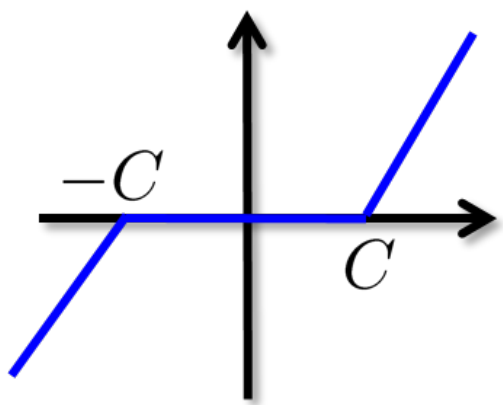
- L_1 regularization: $\psi(x) = C\|x\|_1$.

$$x_{t,j} = \mathrm{ST}_{C\eta_t}(x_{t-1,j} - \eta_t g_{t,j}) \quad (j\text{-th component})$$

where

$$\mathrm{ST}_C(q) = \mathrm{sign}(q)\max\{|q| - C, 0\}.$$

$\rightarrow$ Unimportant elements are forced to be 0.



For many practically used regularizations, analytic form is obtained.

Example of proximal mapping (cont.)

- Trace norm: $\psi(X) = C\|X\|_{\mathrm{tr}} = C\sum_j \sigma_j(X)$ (sum of singular values). Let

$$X_{t-1} - \eta_t G_t = U\,\mathrm{diag}(\sigma_1, \ldots, \sigma_d)V,$$

then

$$X_t = U \begin{pmatrix} \mathrm{ST}_{C\eta_t}(\sigma_1) & & \\ & \ddots & \\ & & \mathrm{ST}_{C\eta}(\sigma_d) \end{pmatrix} V.$$

Convergence of proximal gradient descent

Strong convexity and smoothness of f determines the convergence rate.

$$x_t = \mathrm{prox}(x_{t-1} - \eta_t g_t | \eta_t \psi(x)).$$

property of f	μ-Strongly convex	non-strongly conv
γ-Smooth	$\exp\left(-t\frac{\mu}{\gamma}\right)$	$\frac{\gamma}{t}$
Non-smooth	$\frac{1}{\mu t}$	$\frac{1}{\sqrt{t}}$

- The step size η_t should be appropriately chosen.

Setting of η_t	Strongly conv	non-strongly conv
Smooth	$\frac{1}{\gamma}$	$\frac{1}{\gamma}$
Non-smooth	$\frac{2}{\mu t}$	$\frac{1}{\sqrt{t}}$

- To achieve this convergence rate, we need to take an average of $\{x_t\}_t$ appropriately; Polyak-Ruppert averaging, polynomially decaying averaging.

Convergence of proximal gradient descent

Strong convexity and smoothness of f determines the convergence rate.

$$x_t = \mathrm{prox}(x_{t-1} - \eta_t g_t | \eta_t \psi(x)).$$

property of f	μ-Strongly convex	non-strongly conv
γ-Smooth	$\exp\left(-t\sqrt{\frac{\mu}{\gamma}}\right)$	$\frac{\gamma}{t^2}$
Non-smooth	$\frac{1}{\mu t}$	$\frac{1}{\sqrt{t}}$

- The step size η_t should be appropriately chosen.

Setting of η_t	Strongly conv	non-strongly conv
Smooth	$\frac{1}{\gamma}$	$\frac{1}{\gamma}$
Non-smooth	$\frac{2}{\mu t}$	$\frac{1}{\sqrt{t}}$

- To achieve this convergence rate, we need to take an average of $\{x_t\}_t$ appropriately; Polyak-Ruppert averaging, polynomially decaying averaging.
- Convergence for smooth loss can be improved by Nesterov's acceleration.
  $\rightarrow$ Optimal rate

Outline

1. First order method
   - Proximal gradient descent
   - Nesterov's acceleration and optimal convergence

Nesterov's acceleration (non-strongly convex)

$\min_x\{f(x)+\psi(x)\}$

Assumption: $f(x)$ is γ-smooth.

Nesterov's acceleration scheme

Let $s_1 = 1$ and $\eta = \frac{1}{\gamma}$, and iterate the following for $t = 1, 2, \ldots$

1. Let $g_t \in \partial f(y_t)$, and update $\quad x_t = \mathrm{prox}(y_t - \eta g_t | \eta\psi)$.
2. Set $\quad s_{t+1} = \frac{1+\sqrt{1+4s_t^2}}{2}$.
3. Update $\quad y_{t+1} = x_t + \left(\frac{s_t - 1}{s_{t+1}}\right)(x_t - x_{t-1})$.

If f is γ-smooth, then

$$f(x_t) - f(x^*) \le \frac{2\gamma\|x_t - x^*\|}{t^2}.$$

- This is also called Fast Iterative Shrinkage Thresholding Algorithm (FISTA) (**?**).
- The step size $\eta = 1/\gamma$ can be adaptively determined: back-tracking.
- "Momentum" method is important for deep learning (**?**).

Nesterov's acceleration (strongly convex)

$\min_x \{f(x) + \psi(x)\}$

Assumption: $f(x)$ is γ-smooth and μ-strongly convex. (it must be $\gamma > \mu$)

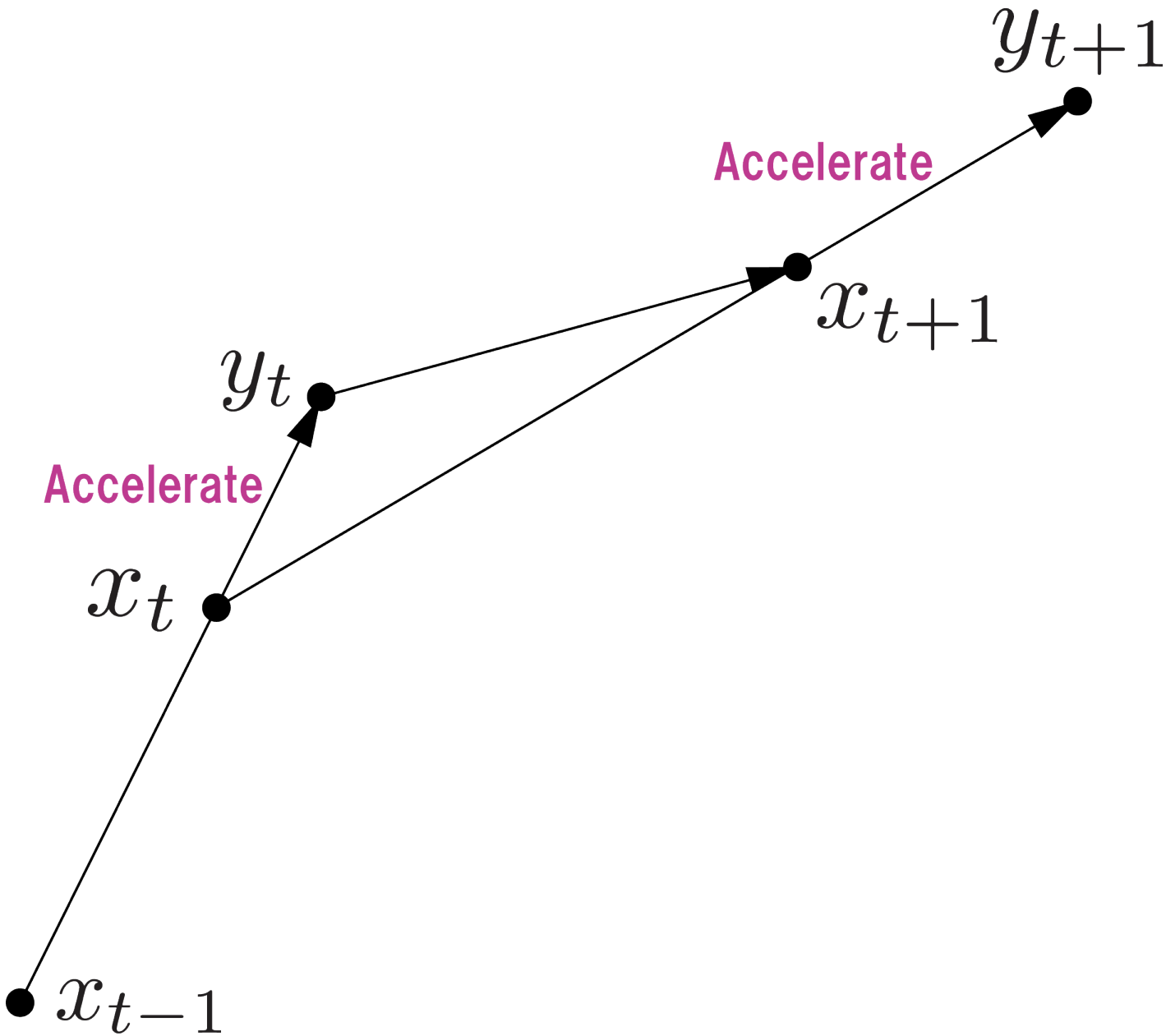
Nesterov's acceleration scheme

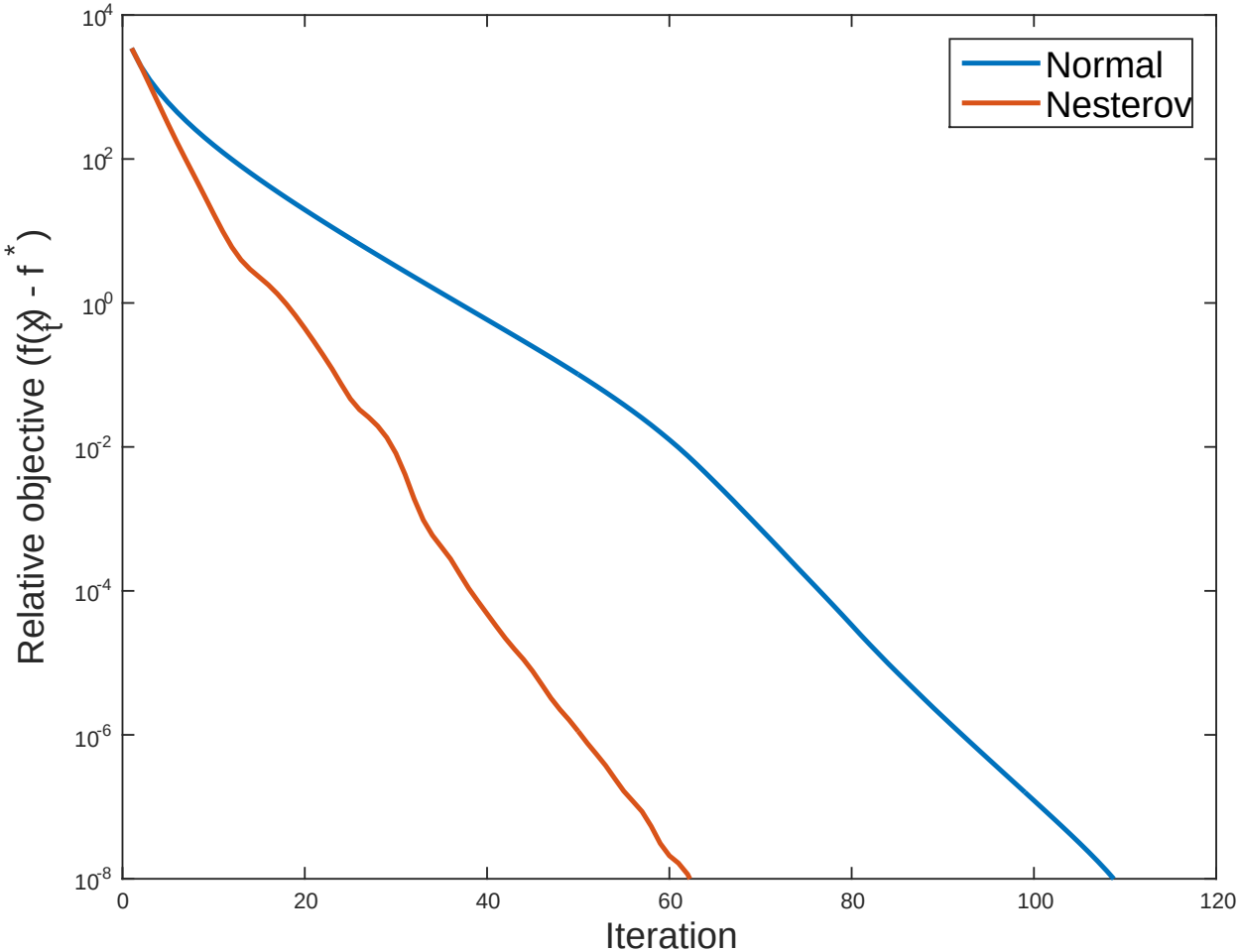
Let $A_1 = 1$, $\alpha_1 = \gamma/\mu$ and $\eta = \frac{1}{\gamma}$, and iterate the following for $t = 1, 2, \ldots$

1. Let $g_t \in \partial f(y_t)$, and update $x_t = \mathrm{prox}(y_t - \eta g_t | \eta\psi)$.
2. Set $\alpha_{t+1} > 1$ so that $(\gamma - \mu)\alpha_{t+1}^2 - (2\gamma + A_t)\alpha_{t+1} + \gamma = 0$, and let $A_{t+1} = A_t/\alpha_{t+1}$.
3. Update $y_{t+1} = x_t + \left(\frac{\mu + A_t}{(\gamma-\mu)(\alpha_{t+1}-1)(\alpha_t - 1)}\right)(x_t - x_{t-1})$.

If f is γ-smooth and μ-strongly convex, then

$$f(x_t) - f(x^*) \le \gamma \left(1 - \sqrt{\frac{\gamma}{\mu}}\right)^t \|x_0 - x^*\|^2.$$


y_{t+1}
Accelerate
x_{t+1}
y_t
Accelerate
x_t
x_{t-1}



Nesterov's acceleration v.s. normal gradient descent

Lasso: $n = 8,000$, $p = 500$.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM journal on imaging sciences, 2(1): 183–202, 2009.

O. Güler. On the convergence of the proximal point algorithm for convex minimization. SIAM Journal on Control and Optimization, 29(2): 403–419, 1991.

I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In Proceedings of the 30th international conference on machine learning (ICML-13), pages 1139–1147, 2013.